# Konferencja Why R?

Warszawa, 27-29 września 2017

# Contents

# Komitet organizacyjny

**Marcin Kosiński (Przewodniczący)**,

Olga Sulima, *Appsilon*

Przemysław Biecek, *Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski*, *Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska*

Bartosz Sękiewicz, *HTA Consulting*

Maciej Beręsewicz, *Uniwersytet Ekonomiczny w Poznaniu*

Adolfo Alvarez, *Uniwersytet Adama Mickiewicza w Poznaniu*, *University of Cincinnati*

Alicja Gosiewska, *Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska*

Aleksandra Dąbrowska, *Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski*

Monika Stępień, *Wydział Biologii, Uniwersytet Warszawski*

Agnieszka Dumania, *SalesBI*

Anna Rybińska, *Uniwersytet Gdański*

Bartłomiej Tartanus, *OSA/Sages*

Krzysztof Słomczyński,

Emil Buszyło, *Maxus Net Communication*

Konrad Więcko, *Samsung*

# Wstęp

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed convallis ut magna at egestas. Vestibulum feugiat lobortis leo ut accumsan. Phasellus sed mattis nibh. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nulla ultrices leo et turpis facilisis ullamcorper. In in augue scelerisque, aliquam magna id, laoreet lacus. Nam pulvinar ipsum vel diam sodales, nec gravida dui ultrices. Donec finibus sit amet ex ac dictum. Suspendisse potenti. Nam varius turpis et tincidunt iaculis. Aenean vel mauris eget diam dictum sodales ac quis arcu. Donec in augue id nisi mattis ultrices.

Cras porttitor, tellus vel auctor dignissim, diam elit bibendum nunc, at tempor ipsum leo vitae elit. Quisque tincidunt nisi sapien, eget pretium ipsum scelerisque in. Mauris porttitor, lacus vitae molestie placerat, justo elit consectetur ante, non bibendum mi tortor eu tellus. Proin luctus elit mi, vel suscipit sem feugiat egestas. Nunc at nisl id ex blandit finibus viverra a libero. Cras faucibus libero in nulla eleifend, sit amet iaculis sapien congue. Aliquam id lacus vel ex posuere maximus sed a massa. Nullam sed lacus at erat tristique dictum. Curabitur consectetur ligula tortor, eget varius massa vestibulum a. Suspendisse potenti.

Suspendisse hendrerit aliquam laoreet. Nunc tempor, ligula sit amet consequat convallis, odio mi tempor justo, varius sodales metus sem vel arcu. Cras nec felis urna. Etiam egestas sagittis tellus, a egestas felis lacinia ut. Proin odio velit, ullamcorper lacinia condimentum a, bibendum vitae enim. Proin feugiat gravida turpis, cursus suscipit risus vehicula a. Donec posuere eros ipsum, in venenatis lectus imperdiet eget. Nam vitae ullamcorper nisl. Cras luctus id diam pretium fringilla. Etiam eleifend urna ac erat laoreet, eget scelerisque nunc hendrerit. Aenean ac vulputate risus. Mauris id rutrum mauris. Ut viverra a nisi ac cursus. Vestibulum auctor, tortor et tempor malesuada, libero tellus egestas nulla, id facilisis nulla massa in ipsum. Praesent iaculis porttitor velit, eget eleifend turpis sollicitudin quis. Mauris volutpat metus sodales porttitor tempor.

# Plan

# Wykłady plenarne

# Show Me Your Model


## Przemysław Biecek

MI$^2$

Kontakt: przemyslaw.biecek@gmail.com


Gramatyka grafiki (Wilkinson, Leland. 2006. The Grammar of Graphics) i jej implementacje (Wickham, Hadley. 2009. Ggplot2: Elegant Graphics for Data Analysis) zmieniły sposób w jaki myślimy o wizualizacji danych. Podobna rewolucja czeka wizualizacje modeli statystycznych. Podczas referatu przedstawię różne istniejące narzędzia do prezentacji modeli statystycznych (rms, forestmodel and regtools, survminer, ggRandomForests, factoextra, factorMerger) oraz zderzę je z jednolitym podejściem do przetwarzania modeli prezentowanym przez pakiet broom (Robinson, David. 2017. Broom: Convert Statistical Analysis Objects into Tidy Data Frames). Prezentacje zakończy zbiór doświadczeń dotyczących wizualizacji struktury modelu (Wickham, Hadley, Dianne Cook, and Heike Hofmann. 2015. Visualizing Statistical Models: Removing the Blindfold. Statistical Analysis).

# Sorted L-One Penalized Estimation


## Małgorzata Bogdan

Uniwersytet Wrocławski

Kontakt: Malgorzata.Bogdan@uwr.edu.pl



SLOPE (Sorted L-One Penalized Estimation) to nowy algorytm optymalizacji wypuklej, sluzacy do redukcji wymiaru w duzych bazach danych. W czasie wykladu zaprezentujemy zastosowanie SLOPE do szeregu problemow statystycznych, jak np. wybor istotnych zmiennych w regresji liniowej i logistycznej czy optymalizacja portfela, a takze omowimy odpowiednie pakiety w R.

# Metody wizualizacji danych jakościowych w programie R


## Justyna Brzezińska

Uniwersytet Ekonomiczny w Katowicach

Kontakt: justyna_brzezinska@ue.katowice.pl



W referacie zaprezentowane zostaną metody wizualizacji danych jakościowych z użyciem odpowiednich pakietów programu R. Przedstawione zostaną podstawowe metody analizy danych jakościowych zapisanych w postaci dwu- i wielowymiarowych tablic kontyngencji, modele przeznaczone do analizy liczebności w tablicach kontyngencji, a także nowoczesne metody i techniki wizualizacji danych o charakterze niemetrycznym. W referacie przedstawione zostaną takie wykresy jak: wykres mozaikowy, sitkowy, asocjacji, dwuwarstwowy, czteropolowy, czy też cpcp oraz rmp.

# Dlaczego czasem R? Why (sometimes) R?


## Tomasz Burzykowski

Hasselt University

Kontakt: tomasz.burzykowski@uhasselt.be



W prezentacji przedstawiona zostanie odpowiedź na pytanie zawarte w tytule wystąpienia. Odpowiedź udzielona zostanie z punktu widzenia biostatystyka zajmującego się od ponad 30 lat analizą danych klinicznych, jak również organizacją prób klinicznych we współpracy z jednostkami akademickimi i firmami farmaceutycznymi.

# Kim naprawdę był Gall Anonim? Zagadnienia statystycznej analizy tekstu


## Maciej Eder

Instytut Języka Polskiego PAN

Kontakt: maciejeder@gmail.com



Wystąpienie będzie poświęcone analizie tekstu za pomocą kilku pakietów języka R, w tym atrybucji autorskiej opartej o statystyczne miary podobieństwa tesktów, a także szeroko rozumianej analizy stylu. Jako jeden z przykładów zostanie omówiony przykład autorstwa "Kroniki polskiej", przypisywanej tzw. Gallowi Anonimowi. W dalszej części wystąpienia zostanie przedstawiona metoda modelowania tematycznego (topic modeling) i jej zastosowania w analizie tekstu.

# R jako główna platforma do zaawansowanej analityki w Enterprise


## Wit Jakuczun

WLOG Solutions

Kontakt: wit.jakuczun@wlogsolutions.com



Świat hermetycznych platform analitycznych powoli staje się historią. Dzisiaj analityka zaawansowana jest pchana do przodu przez świat open-source wspierany przez największych graczy. W różnych dyskusjach stawiane jest pytanie o dojrzałość R z punktu widzenia wymagań korporacji. Na podstawie wdrażania R w dużym telekomie opowiem dlaczego uważam, że R może być numerem jeden jeśli chodzi o zaawansowaną analitykę w każdej dużej korporacji. Pokażę na co zwrócić uwagę i jakie są plusy i minusy przejścia na R.

# Dezagregacja danych przedziałowych


## Michał Ramsza

Szkoła Główna Handlowa w Warszawie

Kontakt: michal.ramsza@gmail.com



Zostanie przedstawiona metoda dezagregacji danych przedziałowych oraz jej implementacja w języku R.

# Jak dużo mocy (i po co) można wycisnąć z modelu predykcyjnego?


## Artur Suchwałko

QuantUp

Kontakt: artur@quantup.pl


W modelowaniu predykcyjnym często wybieramy bardzo złożone podejścia i modele. Z drugiej strony, często też stosuje się podejścia do modelowania, które wręcz jest wstyd stosować w dzisiejszych czasach.

Predykcję można poprawić na różne sposoby, na przykład poprzez wykorzystanie bardziej złożonych modeli, staranny dobór hiperparametrów, uwzględnienie kosztów błędnej klasyfikacji czy zmianę kryterium optymalizacji.

Pokażę na przykładzie, co to daje dla biznesu oraz jak to zrobić w R.

# BioinfoRmatyka nowotworów


## Ewa Szczurek

Uniwersytet Warszawski

Kontakt: szczurek@mimuw.edu.pl



Rak to choroba genomu. DNA komórek rakowych charakteryzują liczne alteracje, o przytłaczającej złożoności i różnrodności. W referacie przedstawię trzy podejścia analizy tak skomplikowanych danych genomicznych z komórek nowotworowych i wyciągania z nich konkretnych wniosków o mechanizmach tej choroby. Wszystkie trzy: muex, SurvLRT i lem, zostały zaimplementowane w R.

# Analiza danych obserwacyjnych zwiedzających wystawy w Centrum Nauki Kopernika


## Anna Wróblewska

MiNI, Politechnika Warszawska

Kontakt: awroble@gmail.com



W prezentacji podsumowujemy współpracę z Centrum Nauki Kopernik (CNK) (Katarzyna Potęga), Uniwersytetem Nauk Społecznych i Humanistycznych (Łukasz Tanaś) oraz Wydziałem Matematyki i Nauk Informacyjnych Politechniki Warszawskiej (WUT). Pracowaliśmy nad danymi obserwacyjnymi zebranymi podczas testów przeprowadzonych w CNK. Dane te zostały przeanalizowane przez studentów nowej specjalności Przetwarzanie i analiza danych na wydziale MINI PW. Dane zostały zebrane z trzech badań obserwacyjnych dotyczących zachowania dzieci i rodziców oraz dzieci szkolnych, a także testów dotyczących rozpoznawania i zrozumienia emocji. Postawiliśmy wiele hipotez i pytań badawczych, zweryfikowaliśmy je w oparciu o dostępne dane, np. podsumujemy różne postawy rodziców, a także zaufanie i rozpoznanie emocji oraz zaangażowanie dzieci w oglądane/doświadczane eksponaty.

# Sesje

# Tidyverse

### Jak zbudowaliśmy aplikację Shiny dedykowaną 700 użytkownikom?

**Olga Mierzwa**

Appsilon

Shiny jako technologia udowodniła, że jest świetnym narzędziem za pomocą, którego zespoły data science mogą komunikować swoje rezultaty. Jednak stworzenie aplikacji w Shiny, która będzie wykorzystywana przez dziesiątki użytkowników nie jest prostym zadaniem. Pierwsze wyzwanie to stworzenie interfejsu użytkownika, który swoim wyglądem nie odbiega od współczesnych rozwiązań. Następnie aplikacja powinna działać wydajnie, co nie raz jest trudne do zapewniania, gdy wzrasta zarówno logika biznesowa i liczba użytkowników.

Celem tej prezentacji jest podzielenie się doświadczeniami jakie nasz zespół data science zdobył budując aplikację obecnie wykorzystywaną przez 700 użytkowników. Skala aplikacji jest jednym z największym produkcyjnych wdrożeń Shiny-ego.

Przedstawimy innowacyjne podejście do tworzenia pięknych i nowoczesnych interfejsów użytkownika za pomocą biblioteki shiny.semantic (alternatywy do obecnego Bootstrapa). Kolejno pokażemy triki, za pomocą których optymalizowaliśmy wydajność aplikacji. Omówimy wyzwania i przedstawimy rozwiązania w zarządzaniu skomplikowanymi zależnościami zmiennych reaktywnych. Zademonstrujemy aplikację i powiemy jak jej wdrożenie przełożyło się na biznes klienta.

### shiny.collections: Google Docs-like live collaboration in Shiny

**Marek Rogala**

Appsilon Data Science

What users expect from web applications today differs dramatically from what was available 5 years ago. They are used to interactivity, data persistence, and what's more, the ability to share live collaboration experiences, like in Google Docs. If one user changes the data, other users want to see the changes immediately on their screens. They don't care whether it is a data-exploration app from a data scientist or a solution built by a team of software engineers.

Shiny is perfect for building interactive data-driven applications suited for the modern user. In this presentation, we show how to create real-time collaboration experience in Shiny apps.

From the presentation, you will learn the concepts of reactive databases, how to use them in Shiny, and how to adapt existing components to provide live collaboration.

We will present a package we developed for that. shiny.collections adds persistent reactive collections that can be effortlessly integrated with components like Shiny inputs, DT::dataTable or rhandsontable. The package makes it easy to build collaborative Shiny applications with persistent data.

The presentation will be very actionable. My goal is for everyone in the audience to be able to add persistence and collaboration to their apps in under 10 minutes.

## Zastosowanie pakietu Shiny do tworzenia interaktywnych wizualizacji wyników badań eye-trackingowych

**Marek Młodożeniec**

OPI PIB

Największym ograniczeniem popularnych metod wizualizacji wyników badań eye-trackingowych, jakimi są wykresy typu 'scanpath' i 'heatmap', jest ich statyczność. Zwłaszcza te ostatnie nie odzwierciedlają w ogóle funkcji czasu, a jedynie przestrzenne rozłożenie fiksacji, natomiast na wykresach typu 'scanpath' przebieg sakad i fiksacji jest często trudny do prześledzenia. Z kolei tworzenie animacji ukazujących przebieg uwagi wzrokowej wymaga zastosowania specjalistycznych programów, w których wygenerowanie animacji wprost z danych surowych bywa problematyczne. Pakiet R Shiny umożliwia tworzenie w prosty sposób interaktywnych aplikacji, które nie tylko pozwalają na śledzenie przebiegu procesu uwagi wzrokowej w formie animacji oraz przewijanie w czasie historii przeszukiwania wzrokowego, ale również dają możliwość regulowania parametrów graficznych wizualizacji, tak aby zapewnić jej maksymalną czytelność. Na przykładzie aplikacji R Shiny pokażę, że w zastosowaniu do wizualizacji danych eye-trackingowych pakiet ten tworzy nową jakość, pozwalając na zaprezentowanie odbiorcom informacji trudnych do ukazania na wykresach innego typu.

### Blog z RMarkdownem i Jekyllem

**Natalia Potocka**

Grupa Wirtualna Polska

Dzięki pakietowi RMarkdown tworzenie stron internetowych jest naprawdę proste. W czasie prezentacji pokażę w jaki sposób pisać bloga lub prowadzić inną stronę internetową mając za narzędzie jedynie RStudio (i parę pakietów). Zaprezentuję proces tworzenia tego typu strony od A do Z oraz wskażę plusy i minusy takiego sposobu utrzymywania strony.

### 10 trików dla wizualizacji w ggplot2

**Piotr Sobczyk**

Infermedica

Wszyscy chcemy tworzyć piękne wizualizacje i za pomocą R staje się to możliwe! Opowiem o niepodstawowych zastosowaniach ggplot2, który jest najbardziej popularnym pakietem do tworzenia grafiki w R. Jego zaletą jest to, że wystarczą jedynie dwie komendy aby stworzyć przyzwoicie wyglądający wykres. Co jednak gdy chcemy zrobić coś zaawansowanego? Podczas prezentacji przedstawię 10 trików na to jak ujarzmić ggplota. Wszystko na podstawie doświadczeń przy tworzeniu bloga szychtawdanych.pl

# PwC Data Analytics

# Statystyka

### Modelowanie dynamiki rozkładu w R. Zastosowanie do analizy konwergencji na poziomie lokalnym

**Piotr Wójcik**

Uniwersytet Warszawski, Wydział Nauk Ekonomicznych

W analizie różnych zjawisk społeczno-ekonomicznych (np. dochodu, osiągnięć edukacyjnych, stopy bezrobocia, preferencji politycznych, wielkości spożycia lodów itp.) często interesujące dla badacza jest ich zróżnicowanie w analizowanej próbie i zmiany tego zróżnicowania w czasie – patrz np. Magrini (2009). Najprostsze podejście ogranicza się do policzenia wybranej miary rozproszenia (np. współczynnika zmienności, współczynnika Giniego, Theila itp.) i porównania jego wartości w kolejnych okresach.

Jednak pojedynczy wskaźnik nie mówi nic o zróżnicowaniu wewnątrz rozkładu. Dlatego inne popularne podejście bierze pod uwagę pełen rozkład badanego zjawiska i polega na porównywaniu histogramów albo jednowymiarowych estymatorów jądrowych w kolejnych okresach. To jednak wciąż nie mówi nic o mobilności wewnątrz rozkładu i nie pozwala formułować długookresowych przewidywań (rozkłady ergodyczne).

Jest to możliwe kiedy dynamika rozkładu jest modelowana za pomocą macierzy przejścia (co wymaga dyskretyzacji rozkładu) albo estymatorów warunkowej funkcji gęstości po raz pierwszy zaproponowanych przez Quaha (1996). Celem prezentacji jest pokazanie jak różne podejścia do modelowania dynamiki rozkładu mogą być zastosowane w R, ze szczególnym uwzględnieniem macierzy przejścia i estymacji jądrowej. Zaprezentujemy zastosowanie w R metodologii umożliwiającej podsumowanie dwuwymiarowego warunkowego estymatora gęstości za pomocą jednowymiarowego rozkładu ergodycznego – patrz Gerolimetto i Magrini (2017).

Przedstawimy także czytelne i atrakcyjne sposoby wizualizacji wyników estymacji. Praktyczne przykłady dotyczące modelowania procesów lokalnej konwergencji będą oparte na danych symulowanych oraz na rzeczywistych danych przestrzennych.

Literatura Gerolimetto, Margherita, and Stefano Magrini. 2017. “A Novel Look at Long-Run Convergence Dynamics in the United States.” International Regional Science Review 40 (3). Magrini, Stefano. 2009. “Why Should We Analyse Convergence Through the Distribution Dynamics Approach?” Science Regionali 8: 5–34. Quah, Danny. 1996. “Twin Peaks: Growth and Convergence in Models Distribution Dynamics.” Economic Journal 106: 1045–55. Silverman, B.W. 1986. Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. Londyn: Chapman; Hall.

## Relacje między lasem a klimatem w różnych skalach przestrzennych – jak to ugRyźć?

**Marcin K. Dyderski Andrzej M. Jagodziński**

Instytut Dendrologii Polskiej Akademii Nauk w Kórniku

Lasy poprzez wiązanie dwutlenku węgla z atmosfery regulują jego stężenie, wpływając na klimat. Z drugiej strony, warunki klimatyczne wyznaczają granice występowania poszczególnych gatunków drzew. Celem prezentacji jest pokazanie przykładów zastosowania bibliotek programu R jako narzędzi które pomagają poznać oba procesy w różnych skalach przestrzennych. Wiązanie dwutlenku węgla przez drzewa związane jest z produkcją biomasy. Pakiet dplR pozwala na analizy sekwencji przyrostów rocznych drzew – szeregów czasowych, skorelowanych głównie z warunkami klimatycznymi oraz wiekiem drzewa. Dzięki znajomości zmian średnicy drzewa i allometrycznych zależności pomiędzy średnicą i masą, możemy wykonać predykcję przyrostu biomasy drzew. Znając zawartość węgla w biomasie drzew możemy obliczyć, jak wiele dwutlenku węgla jest pochłaniane przez drzewa. Do obliczenia biomasy w drzewostanach wykorzystuje się wskaźniki przeliczeniowe – tzw. Biomass Conversion and Expansion Factors (BCEF). Dzięki zastosowaniu odpowiednich BCEF oraz danych z inwentaryzacji leśnych można obliczyć zasoby biomasy oraz związanego dwutlenku węgla w lasach Polski. Jednym z rozwiązań ułatwiających obliczenia w tym procesie jest pakiet dplyr, pozwalający na szybką aplikację odpowiednich wskaźników. Czynniki klimatyczne wpływają na występowanie poszczególnych gatunków drzew. W celu określenia zmian zasięgu ich występowania stosuje się modele rozmieszczenia gatunków, w których predyktorami są parametry bioklimatyczne. Po zastosowaniu takiego modelu do projekcji zmian klimatycznych można określić zmiany optimum klimatycznego dla poszczególnych gatunków. Jednym z algorytmów predykcji rozmieszczenia gatunków jest model MaxEnt, zaimplementowany w pakiecie dismo. R jest narzędziem pozwalającym na analizę złożonych danych z zakresu nauk leśnych. Umożliwia to próbę odpowiedzi na najistotniejsze pytania z pogranicza ekologii i hodowli lasu, mające duże znaczenie zarówno naukowe, jak i praktyczne, w obliczu prognozowanych zmian klimatycznych.

## Wydźwięk artykułów prasowych jako narzędzie wspomagające predykcyje kondycji finansowej przedsiębiorstw

**Radomir Nowacki**

Uniwersytet Warszawski

Przedstawienie wyników analizy wydźwięku artykułów z Parkietu z lat 2000-2017(ponad 150 tysięcy artykułów), prezentacja użytych metod i narzędzi, użyteczność tłumaczenia maszynowego w celu skorzystania z bogatszego zestawu bibliotek do języka angielskiego.

## FSelectorRcpp – wolna od Java/Weka implementacja pakietu FSelector

**Krzysztof Słomczyński**

Appsilon

Celem prezentacji będzie zapoznanie użytkowników języka R z procesem powstawania oraz możliwościami pakietu FSelectorRcpp. Zostanie on też zestawiony z innymi popularnymi pakietami służącymi do selekcji zmiennych jak i z jego wcześniejszą – opartą na Java – implementacją.

### Analiza sentymentu przy użyciu bibliotek Microsoft

**Łukasz Grala**

TIDK - Data Scientist as a Service

Analiza sentymentu jest powszechnym wyzwaniem wielu dużych organizacji. Analizujemy treści poczty elektronicznej, dyskusji na forach, tekstów z komunikatorów, czy też różnorodnych portali społecznościowych. Analiz ta jest istotna z punktu widzenia budowania wizerunku firm, produktów, czy też osób. Algorytmów i metod jest wiele, w czasie sesji pokażemy jak to można wykonać dzięki gotowym biblioteką dostarczanym przez firmę Microsoft. Rozwiązanie dostępne zarówno w środowisku Windows, jak i Linux, z poziomu SQL Server, hurtowni danych Teradata, czy też rozwiązań pracujacych na HADOOP czy Spark

## Modelowanie Ryzyka (UBS)

### Modelowanie ryzyka kredytowego z wykorzystaniem panelowej regresji liniowej

**Stanisław Ochotny**

UBS

Typowe modele ryzyka kredytowego (PD) jako zmienną zależną biorą zmienną binarną, wskazującą, czy w danym okresie klient zdołał spełnić swoje zobowiązania finansowe wynikające z umowy kredytowej. Problem pojawia się, gdy w portfelu kredytów

lub segmencie klientów historycznie nie mamy dostatecznie wielu obserwacji wskazujących na niewypełnianie zobowiązań, by zbudować dla niego dobry model statystyczny. Wykład przedstawi inne możliwe zmienne zależne i metody ich modelowania z użyciem panelowej regresji liniowej (pakiet plm).

## Agregacja rozkładów przy pomocy kopul

**Adam Wróbel**

UBS

Celem prezentacji jest zapoznanie uczestników z modelowaniem przy pomocy kopul poprzez przestawienie praktycznego zastosowania.

Sytuacja banku zależy od wielu czynników takich jak stopy procentowe, sytuacja na giełdzie, ceny na rynku nieruchomości. Każdy z tych czynników możemy modelować samodzielnie, ale w czasie kryzysu te czynniki mogą być ze sobą mocno skorelowane. Można to było zaobserwować w czasie ostatniego kryzysu, gdy krach na rynku nieruchomości wywołał kryzys na rynkach finansowych. Dlatego potrzebujemy struktury zależności miedzy czynnikami, aby mieć pełny obraz tego ile bank potrzebuje kapitału, aby przetrwać nawet w skrajnym scenariuszu. Taką strukturę zależności możemy zdefiniować wykorzystując kopule.

Pakiety: CDVine, ghyp, dplyr

## Zastosowanie analizy szeregów czasowych w modelowaniu ryzyka niespłacalności

**Dorota Kowalczyk**

UBS

W następstwie kryzysu finansowego 2007 -2009 banki przywiązują dużą wagę do prognozowania strat dla potrzeb testów stresu. Testy stresu polegają na estymowaniu straty w zadanym scenariuszu makroekonomicznym. Elementem prognozowania strat z tytułu ryzyka kredytowego jest zwykle prognozowanie ryzyka niespłacalności (PD - probability of default). Prognozy takie tworzone są z wykorzystaniem makroekonomicznych czy też finansowych szeregów czasowych. Po krótkim wprowadzeniu do modelowaniu ryzyka niespłacalności dla potrzeb testów stresu i do wybranych elementów analizy szeregów czasowych (np stacjonarność czy rząd integracji), skupimy się pakietach zwykle wykorzystywanych do takiej analizy: uroot, urca , tseries. Niektóre z zastosowanych w tych

pakietach rozwiązań zostaną poddane krytyce, inne opatrzone komentarzem jak je lepiej stosować.

Pakiety: uroot, urca , tseries

# Biostatystyka

### Identyfikacja i analiza barier przeciwko introgresji pomiedzy gatunkami roślin z rodzaju Capsella (Tasznik)

**Krzysztof Stankiewicz**

Imperial College London

Zbudowaliśmy model typu HMM (z ang. Ukryte Pole Markowa) do zidentyfikowania introgresji DNA pomiędzy gatunkami roślin z rodzaju Capsella (Tasznik). Rezultaty z tych analiz były użyte do badania barier przeciwko introgresji, które powodują amplifikację różnic pomiędzy subpopulacjami i następującym rozszczepem gatunków w procesie ewolucyjnym.

### Przegląd pakietów służących do analizy ekspresji genów i metylacji DNA

**Alicja Gosiewska**

Politechnika Warszawska

W swoim wystąpieniu przedstawię pakiet łączący najczęściej używane pakiety z Bioconductora dotyczące ekspresji oraz metylacji. Pokażę również, jak przy jego pomocy można dokonywać wizualizacji wyników otrzymanych przy użyciu testów statystycznych.

### Metylacja DNA i ekspresja genu na przykładzie danych RTCGA.

**Aleksandra Dąbrowska**

Uniwersytet Warszawski

W swojej prezentacji przedstawię wyniki użycia pakietu łączącego metody analizy metylacji DNA i ekspresji genu na podstawie danych z pakietu RTCGA.

## survminer - wykresy analizy przeżycia pełne informacji i elegancji

**Marcin Kosiński**

Data Applications Designer

survminer to pakiet w R, który na scenie analizy przeżycia wypełnia lukę wizualizacji estymatorów krzywych przeżycia w duchu 'Grammar of Graphics' (ggplot2). W trakcie prezentacji przedstawię jak wyjątkowo elastyczne i konfigurowalne jest to narzędzie do tworzenia wykresów krzywych przeżycia. Wyjaśnię także czym są te wykresy oraz jak je interpretować. Warto rozumieć tę metodologię, ponieważ skala zastosowań analizy przeżycia jest rozpięta niemalże nad każdą dziedziną życia - od kontroli jakości żarówek, przez wyliczanie składek ubezpieczeniowych aż do badań klinicznych nad nowotworami. Jeżeli starczy czasu zaprezentuję także funkcjonalności survminer'a do diagnostyki i sprawdzenia założeń modelu Coxa proporcjonalnych hazardów - najbardziej popularnej metody statystycznej w analizie przeżycia, która niekoniecznie jest najlepiej rozumiana.

## Podstawy przetwarzania i analizy obrazów w R

**Andrzej Oleś**

EMBL Heidelberg

W oparciu o pakiet do przetwarzania i analizy obrazów EBImage zademonstrowane zostaną metody pracy z danymi graficznymi w R: wczytywanie i wyświetlanie, transformacje przestrzenne oraz filtrowanie. Na przykładzie mikroskopowych obrazów komórek pokazane zostanie jak przeprowadzić segmentację obrazu w celu wyodrębnienia charakterystyk ilościowych obiektów, stanowiących punkt wyjścia do dalszych analiz statystycznych.

# Biznes

## Wyzwania stawiane przez technologie open source w biznesie

**Mikołaj Olszewski Mikołaj Bogucki**

Pearson

W świecie współczesnej analityki danych coraz więcej firm rezygnuje z komercyjnych narzędzi analitycznych na rzecz oprogramowania open source, czego R jest świetnym przykładem. Prowadzi to nie tylko do redukcji kosztów ale często do rozwoju samej technologii przez firmę, która ją zaadoptowała.

Oprogramowanie open source takie jak R ma jednak pewne wady, np. pakiety nie działają zgodnie z oczekiwaniami, nowe wersje pakietów zmieniają lub usuwają stare funkcje, pakiety które zespół używa w codziennej pracy zostają całkowicie porzucone, zmuszając zespół do przyjęcia innych rozwiązań.

Jako analitycy danych w Pearsonie, wykorzystujemy R i Shiny jako główne narzędzia do przetwarzania danych, wizualizacji i raportowania. W naszej prezentacji przedstawimy faktyczne wyzwania, z którymi przyszło nam się zmierzyć w ciągu ostatnich kilku lat. Przedstawimy rozwiązania, które zastosowaliśmy oraz ich wpływ zarówno na bieżące projekty, jaki i na podejście zespołu do nowych problemów.

## Kiedy data.frame pożera Workfile, czyli o tym jak przeprowadzamy stadko ekonometryków z klikania w EViewsie do pisania w R

**Anna Skrzydło**

MediaCom Warszawa Sp. z o. o.

Zmiana zawsze jest trudna. Każda zmiana. A szczególnie taka, która wymaga przeniesienia się z przyjaznego świata klikalnego oprogramowania do surowego, białego ekranu R Studio. Czy można zamienić kilkunastoosobowy zespół ekonometryków w programistów? Czy może to zajęcie tylko dla wybranych, którzy przywdziewając flanelowe koszule tworzą narzędzia jak najbardziej podobne do znanych i lubianych softów? Krótka opowieść o tym jak przenosimy nasz proces modelowania ekonometrycznego z EViewsa do R, jakie wyzwania stoją na naszej drodze i co dzięki tej zmianie zyskujemy.

## Czy R (kiedyś) zastąpi SPSS?

**Tomasz Zółtak**

Instytut Badań Edukacyjnych

R staje się coraz popularniejszym narzędziem również w dziedzinie nauk społecznych, w tym w analizie danych sondażowych. Ogromne znaczenie dla tego obszaru zastosowań miał rozwój możliwości związanych z tworzeniem raportów, jaki nastąpił w ostatnich latach. Niemniej R wciąż nie jest wymarzonym narzędziem do "robienia tabelek" prezentujących wyniki typowych sondaży. W wystąpieniu zamierzam zarysować specyficzne problemy związane z analizą danych sondażowych oraz zastanowić się, jakie rozwiązania niezbędne do wygodnego prowadzenia takich analiz są już dostępne w środowisku R, a jakich elementów moim zdaniem wciąż brakuje.

## R a dane w chmurze

**Krzysztof Jędrzejewski Emilia Pankowska**

Pearson

W międzynarodowej korporacji sprawne działanie zespołów analitycznych rozsianych po całym świecie wymaga efektywnego udostępniania, współdzielenia oraz przetwarzania danych. Jedną z możliwości osiągnięcia tego celu w prosty sposób jest skorzystanie z usług chmurowych świadczonych przez specjalizujące się w tym firmy. Jednym z najpopularniejszych dostawców rozwiązań w tym obszarze jest firma Amazon. Na przykładzie jednego z naszych projektów analitycznych opiszemy w jaki sposób z poziomu języka R można przetwarzać dane z wykorzystaniem usług amazonowych. Przedstawimy kilka podejść jakie przetestowaliśmy, ze szczególnym uwzględnieniem ich zalet i wad, oraz problemów jakie napotkaliśmy.

## Podejmowanie decyzji w R - w warunkach pewności, ryzyko oraz niepewności

**Vasyl Martenyuk**

Akademia Techniczno-Humanistyczna w Bielsku-Białej

Celem wystąpienia jest przedstawienie zagadnień komputerowego wspomagania pro-

cesów podejmowania decyzji w R. Będą rozpatrywane decyzyjne problemy w warunkach pewności, ryzyko, niepewności oraz przedstawiono odpowiednie narzędzie R. Podejścia będą ilustrowane rzeczywistymi przykładami z branży e-marketingu

# Społeczności

# R w działaniu

### R-Ladies Warsaw

**Natalia Potocka**

R-Ladies Warsaw

### Projekt R w Google Summer of Code: okiem mentora

**Tomasz Melcer**

QuantUp

Google Summer of Code to program umożliwiający sfinansowanie płatnych wakacyjnych praktyk studenckich związanych z tworzeniem oprogramowania open-source. Projekt R bierze udział w tym programie od 2008 roku. Na wystąpieniu opowiem jak wygląda praca w ramach tego programu: od etapu zgłaszania pomysłów na projekty studenckie do końcowych rozliczeń. Na co zwrócić uwagę, jakich problemów można się spodziewać i jak organizować pracę, by staż zakończył się sukcesem.

### Planowanie pojemności i wydajności w celu przeciwdziałania awariom

**Robert Bigos**

Wcześniej pracował dla fim takich jak IBM, SAP, Sabre. Obecnie doradza klientom ciesząc się życiem :)

Nie ma zasobów nieskończonych, nie ma zasobów idealnych. Każdy system komputerowy to zbiór bardzo wielu zależnych od siebie kolejek które mogą zostać przeciążone i wpływać na pozostałe. Przy odpowiednim przeciążeniu wszystko kiedyś ulegnie awarii. Coraz częściej spotykamy się ze zjawiskiem “capacity rolling disaster” czyli awariami w których pierwotna przyczyna tylko inicjuje łańcuch zdarzeń. Duże systemy instrumentacyjne (logi/metryki) zbierają rocznie TBy danych, w rozdzielczość na poziomie (ms)sekund. Jak w tym wszystkim doszukać się wzorców i pierwotnych przyczyn awarii by im przeciwdziałać? Jak zaplanować zmiany aby zapewnić odpowiednią pojemność systemu w czasie?

Do zabawy z danymi wykorzystamy R i wizualizacje grafów animowanych po czasie.

## Jak wygrać więcej w lotto z R?

**Błażej Kochański**

Politechnika Gdańska, Wydział Zarządzania i Ekonomii, Katedra Nauk Ekonomicznych

Dostępne w Internecie dane dotyczące poprzednich losowań oraz wygranych w loterii nazywanej kiedyś "Dużym Lotkiem" (6 z 49) mogą pomóc w ukształtowaniu optymalnej strategii gry. Jeden z kluczy: wybieranie mniej popularnych liczb. Celem jest maksymalizacja wartości oczekiwanej kwoty wygranej. Uzyskanie wartości oczekiwanej większej niż cena losu (R pomoże stwierdzić, czy to możliwe) prowadzi do pytania, jaką część majątku możemy inwestować? Z pomocą przychodzi tzw. kryterium Kelly’ego.

# Lightning Talks

*Sala 107 - Chairman: Marcin Kosiński*

---

# Lasy z inwazyjnym dębem czerwonym w świetle analiz wielowymiarowych przy użyciu R

---

**Damian Chmura**

Akademia Techniczno-Humanistyczna w Bielsku-Białej

Północnoamerykański dąb czerwony Quercus rubra L. zaliczany jest do tzw. gatunków inwazyjnych w naszej florze, tzn. rozprzestrzenia się spontanicznie i wywiera negatywny wpływ na rodzime gatunki roślin występujące głównie w runie. Celem niniejszych badań była analiza wielowymiarowa (analiza skupień, techniki ordynacyjne i analiza funkcjonalna) składu gatunkowego lasów zastępczych z udziałem dębu czerwonego. W latach 2008-2011 wykonano 180 zdjęć fitosocjologicznych w wybranych losowo kompleksach leśnych z udziałem dębu na Wyżynie Śląskiej. Zebrany materiał poddano analizie skupień. Ze względu na to, że miara odległości ma wpływ na końcowy wynik zastosowano funkcję rankindex (pakiet vegan) a jako matrycę danych siedliskowych użyto średnich arytmetycznych i ważonych liczb Ellenberga. W celu ustalenia kierunków zmienności uzyskane jednostki roślinności przeanalizowano technikami ordynacyjnymi (nietendencyjna analiza zgodności, DCA). Wpływ wybranych czynników siedliskowych na zmienność składu gatunkowego oraz pokrycie dębu w poszczególnych warstwach (warstwa drzew, podszyt i runo) sprawdzono przy użyciu kanonicznej analizy korespondencji CCA, analizy redundancji RDA oraz testów permutacyjnych, pakiety: vegan, ade4). Zastosowano analizę indVal (indicator value), aby stwierdzić czy są istotne statystycznie gatunki wskaźnikowe dla wybranych typów lasów (pakiety: labdsv, indicspecies). Wykonano również analizę taksonomiczną i funkcjonalną analizowanych lasów. Wyliczone wskaźniki: bogactwa gatunkowego i różnorodności gatunkowej (wskaźnik Shannona-Wienera i in.) oraz różnorodności funkcjonalnej (pakiet FD: funkcjonalne bogactwo FRic, funkcjonalna jednorodność FEve i funkcjonalna rozbieżność FDiv) porównano

między badanymi lasami. Wyniki analiz wielowymiarowych pozwalają na określenie, na jakich typach siedlisk dąb czerwony częściej dokonuje inwazji lub wcześniej był sadzony, z jakimi gatunkami współwystępuje oraz jaki jest wpływ tego drzewa na rośliny towarzyszące.

# FasteR - kiedy warto przenieść obliczenia na superkomputery?

**Bartosz Czernecki**

Uniwersytet im. A. Mickiewicza w Poznaniu

W prezentacji przedstawiono przykładowe rozwiązania (i ograniczenia) związane z przyspieszaniem kodu R na przykładzie danych meteorologicznych. Omówiono korzyści płynące z unikania pętli, wektoryzacji obliczeń, wykorzystania bardziej wydajnych pakietów a także przepisania kodu R do C(++)/Fortrana i konsekwencji wynikających ze zrównoleglania obliczeń. Przedstawiono także kilka własnych doświadczeń dotyczących przenoszenia obliczeń na superkomputery (HPC) znajdujące się w Poznańskim i Wrocławskim Centrum Superkomputerowo-Sieciowym w ramach dostępnych dla zastosowań naukowych (i komercyjnych) grantów obliczeniowych.

# SparkR - wydajne obliczenia w chmurze

**Włodzimierz Bielski**

ITMAGINATION

Połączenie R z łatwo dostępną mocą obliczeniową w chmurze pozwala na realizowanie nowych, niedostępnych do tej pory scenariuszy. Pokażę jak prosto możemy sięgnąć po tę moc, korzystając z pakietu SparkR.

# Uprzyjemnij sobie generowanie wielu wykresów za pomocą purrr

**Mateusz Otmianowski**

Pearson

Praca analityka wymaga tworzenia wielu wykresów, szczególnie w fazie eksploracji danych. Jest to często żmudne zajęcie, które można jednak usprawnić poprzez wykorzystanie pakietu purrr. Zademonstruję jak używać purrr w kombinacji z plotly do hurtowego generowania wykresów oraz przetrzymywania ich w data frame'ach, co ogranicza nakład pracy oraz sprawia, że kod jest zwięzły i zrozumiały dla pozostałych analityków.

# What drumming taught me about leading a data science team

**Kacper Lodzikowski**

Pearson

This talk provides practical tips on how to lead a data science team by drawing an analogy between the role of a drummer in a band and a team leader. While drummers don't create the main value of a song (the melody) and they rarely 'lead' bands the way lead singers or guitarists do, they are always their band's backbone because they set and keep time for others to follow. Similarly, good data science leaders use best agile practices to set the rhythm of internal processes of working with data. Moreover, they do everything they can to maintain the rhythm of work and, when other team members miss a beat, to improvise accordingly. Finally, they remember their place is at the back of the band, so that others have the freedom to explore the data in whichever way they think creates most value.

# Zawód rodzica a edukacja dziecka - wizualicja wyników badań PISA

**Mateusz Staniak**

Uniwersytet Wrocławski

Badanie PISA sprawdza wiedzę piętnastolatków z kilkudziesięciu krajów w dziedzinach czytania, matematyki i nauk przyrodniczych. Na podstawie aplikacji napisanej pod opieką Przemysława Biecka, dostępnej pod adresem , pokażę, jak pakiety ggplot2 i shiny pozwalają odkryć zależności pomiędzy zawodami rodziców (które m.in. odzwierciedlają ich status społeczny) i wynikami ich dzieci oraz jak te zależności zmieniały się na przestrzeni lat 2006-2015.

# Sesja plakatowa

*Aula Gmachu Fizyki PW*

## Analizy gleboznawcze w R

**Łukasz Pawlik, Pavel Samonil**

Uniwersytet Pedagogiczny, Kraków
Instytut Ekologii Lasu, Brno, Czechy

Wyniki laboratoryjnych analiz chemicznych i fizycznych próbek gleb pobranych z poligonów badawczych w Gorcach, rezerwacie Zofin (Czechy) i stanie Michigan (USA) zostały poddane analizie i wizualizacji w pakiecie statystycznym R. W tym celu wykorzystano następujące pakiety: stats, aqp, corrplot, FSA, ggplot2, vegan, psych.

## Competition and tourism drive trade-off between vegetative and generative reproduction of rare mountain species Carex lachenalii Schkuhr

**Patryk Czortek**

Uniwersytet Warszawski

A result of sexual reproduction is long-distance spread at a meta-population level, whereas vegetative propagation contributes to increase population growth at a local scale. Trade-offs between these two components of reproduction reflect adaptation to the environment

and may be a suitable way to understand threats of rare key-species. Example of such plant may be Carex lachenalii Schkuhr – a small tufted perennial sedge, occurring in extreme-specialized snowbed and acidophilous grasslands vegetation. This arctic-alpine species in the Tatra Mts occurs only in a few isolated sites. In 2016 we examined 96 localities of C. lachenalii. Maximum height and diameter, number of vegetative and generative stems, all vascular plants within 100m2 plots and the distance from the nearest trail was recorded for each sedge tuft, with the aim of determining whether tourism affects trade-off. To determine the role of competition and habitat filtering in shaping species composition of plant communities we calculated components of functional diversity using FD package. We used principal components analysis (PCA) for detecting relationships between species composition and populational traits of C. lachenalii, using vegan::envfit() function. For evaluation the impact of vegetation traits on trade-off we used generalized additive models (GAM) and chose the best model, based on Akaike's Information Criterion (AIC). We found habitat-dependent relationships between all vegetation traits influencing the studied trade-off. Distance from the nearest trail did not influence the studied traits.

# Informacja publiczna trochę bardziej publiczna - dane z liczników rowerów w Warszawie w Shiny

**Monika Pawłowska**

Instytut Biologii Doświadczalnej im. M. Nenckiego PAN

Dzisiejsze miasta zbierają ogromne ilości danych. Większość z nich stanowi informację publiczną, która powinna być udostępniana obywatelkom i obywatelom. Jednak aktywiści i urzędnicy zwykle nie mają ani odpowiednich umiejętności, aby narzędzi, żeby móc z tych danych skorzystać. Z kolei gotowe raporty publikowane są rzadko i odpowiadają tylko na wybrane pytania.

Pochylając się nad tym problem, wzięliśmy pod lupę dane z automatycznych liczników, które od kilku lat zliczają ruch rowerowy w kilkunastu miejscach w Warszawie. Przy użyciu pakietu Shiny przygotowaliśmy aplikację dostępną pod adresem: http://greenelephant.pl/rowery. Pokazuje ona między innymi natężenie ruchu rowerowego w poszczególnych lokalizacjach w różnym czasie. Można z niej odczytać zależność liczby rowerów od temperatury powietrza i opadów. Dostępna też jest interaktywna mapa lokalizacji liczników.

Shiny pozwala na przedstawienie danych w sposób przystępny, i elastyczny. Umożliwia ekspolarcję danych przez użytkowników bez wiedzy z dziedziny programowania i statystyki. Natomiast dzięki otwartemu kodowi aplikacji może być użyta jako przykład i łatwo zmodyfikowana na potrzeby wizualizacji danych z innych miast lub odmiennego charakteru.

# Mood of Music - a machine learning model for classifying mood of music by evaluating characteristic words of songs lyrics

**Patrycja Cieślak**

Uniwersytet Gdański

poster, omówienie eksperymentu.

# TBA

**Sylwia Wierzcholska**

Uniwersytet Wrocławski

TBA

# Warsztaty

# Analiza danych sondażowych w R

**Dariusz Szklarczyk, Agnieszka Otręba-Szklarczyk**

Fundacja Rozwoju Badań Społecznych

## Opis warsztatu

Wśród badaczy społecznych korzystających używających R do pracy z danymi panuje dość powszechna zgoda, że dużo łatwiej jest w R zbudować np. model regresji liniowej, niż przygotować dane do analizy i wykonać najprostsze, a najpowszechniej stosowane w praktyce analizy, np. rozkłady procentowe dla wielokrotnych odpowiedzi czy tabele krzyżowe. Tymczasem, ze względu na dużą liczbę zmiennych kategorialnych stosowanych w badaniach społecznych, zwłaszcza sondażowych, umiejętność ta jest niezbędna zarówno na etapie wstępnej eksploracji danych, jak i prezentowania prostych zależności. Jednocześnie trudności w przeprowadzeniu prostych, a niezbędnych operacji, skutecznie zniechęcają (niesłusznie!) do nauki R osoby, które przyzwyczaiły się do prostego i błyskawicznego ich sporządzaniu przy pomocy komercyjnych pakietów, takich jak SPSS czy Statistica. Celem warsztatu jest zaprezentowanie najbardziej przydatnych funkcji i pakietów służących przygotowaniu danych z badań sondażowych do analizy (m.in. pakiet dplyr) i analizie danych sondażowych, ze szczególnym uwzględnieniem danych kategorialnych i zależności między nimi (m.in. rozkłady procentowe, wielokrotne odpowiedzi, tabele krzyżowe). Nacisk zostanie również położony na przygotowanie planu analizy. Aby zmaksymalizować użyteczność R w kontekście badań sondażowych, zaprezentowane zostaną również pakiety i funkcje służące do doboru próby (m.in. sampling) oraz ważenia danych sondażowych (weights). Warsztat adresowany jest w szczególności dla badaczy społecznych, politologów, socjologów, badaczy rynku i innych osób, które w pracy zawodowej analizują dane sondażowe i chcieliby zacząć robić to w R, z naciskiem na analizę refleksyjną, niezautomatyzowaną.

## Plan warsztatu

1. Przygotowanie danych do analizy – czyszczenie bazy danych, rekodowanie danych.
2. Eksploracja danych sondażowych – przygotowanie planu analizy, zestawy wielokrotnych odpowiedzi, statystyki opisowe, wizualizacja danych.
3. Analiza zależności między danymi sondażowymi – tabele krzyżowe, analiza korespondencji, analiza zależności między zmiennymi ilościowymi, tworzenie indeksów.
4. Dobór próby losowej (prostej, warstwowanej) i ważenie próby (klasyczne oraz rake weights).

**Wymagane pakiety**

dplyr, sampling, weights, questionr, ca, ggplot, gmodels

**Wymagane od uczestników umiejętności i wiedza**

Zapraszamy wszystkie osoby zainteresowane tematem analizy danych sondażowych, społecznych. Mile widziane podstawowe doświadczenie w prowadzeniu badań sondażowych, ankiet itp.

**Wymagania wstępne do wykonania przed warsztatem**

Zainstalowanie R, RStudio. Dane zostaną przekazane w trakcie warsztatu.

# Złożone schematy doboru próby - pakiet survey

**Tomasz Żółtak**

Instytut Badań Edukacyjnych

## Opis warsztatu

Duża część dostępnych powszechnie danych z badań sondażowych pochodzi z projektów, w których wykorzystywane są złożone schematy doboru prób badawczych. W szczególności dotyczy to międzynarodowych badań porównawczych w dziedzinie edukacji (np. TIMSS, PIRLS, PISA, PIAAC) czy nauk społecznych (np. ESS), ale też badań dotyczących zdrowotności i epidemiologii. Analiza tych danych przy pomocy klasycznie wykorzystywanych technik, zakładających, że próba została dobrana w sposób prosty, może prowadzić do błędnych wniosków, w szczególności w zakresie wielkości błędów standardowych (a w konsekwencji istotności statystycznych). Zasadniczym celem warsztatu jest zapoznanie uczestników z możliwościami pakietu „survey", który umożliwia analizę tego rodzaju danych w R z wykorzystaniem technik adekwatnych do prób dobranych w sposób złożony: z wykorzystaniem stratyfikacji, doboru wielostopniowego, czy zespołowego.

## Plan warsztatu

1. Złożone schematy doboru prób badawczych – jak i po co się to robi?
   (a) Typowe złożone schematy doboru próby: stratyfikacja, dobór zespołowy i wielostopniowy.
   (b) Przykłady projektów badawczych, w których wykorzystywane są złożone schematy doboru próby.
   (c) Estymacja wariancji estymatorów z wykorzystaniem linearyzacji Taylora i z wykorzystaniem technik replikacyjnych: podstawowe pojęcia i założenia oraz ich najważniejsze implikacje praktyczne.
2. Pakiet „survey" - jego możliwości i ograniczenia.
   (a) Co możemy zrobić z pakietem „survey".
   (b) Inne możliwości ale w specyficznych zastosowaniach: pakiety „intsvy" i „lavaan.survey".
3. Definiowanie typowych złożonych schematów doboru próby w pakiecie „survey".
4. Estymacja typowych statystyk opisowych.
   (a) Średnie, wariancje, kwantyle (i sumy populacyjne!).
5. Przerwa.
6. Wizualizacja danych przy pomocy pakietów „survey" i „ggplot2".

(a) Funkcje graficzne pakietu „survey”.

(b) Wykorzystywanie wag w pakiecie „ggplot2”.

7. Regresja liniowa i uogólniona regresja liniowa.
   (a) Jak korzystać z funkcji ‘svyglm()’?
   (b) A co z liczeniem korelacji?
   (c) W jakich sytuacjach warto uwzględniać złożony schemat doboru próby przy prowadzeniu analizy regresji?
8. Poststratyfikacja i techniki pokrewne (jeśli ktoś jest zainteresowany, może rzucić okiem na tą prezentację).
   (a) Co to znaczy, że próba jest reprezentatywna? (Bardzo możliwe, że nie to, czego się spodziewasz!)
   (b) Definiowanie wag poststratyfikacyjnych w pakiecie „survey”.
   (c) Kiedy warto używać poststratyfikacji, a kiedy lepiej tego nie robić?

W czasie warsztatu wykorzystywane będą dane z Europejskiego Sondażu Społecznego i badań PISA.

## Wymagane pakiety

survey, ggplot2

## Wymagane od uczestników umiejętności i wiedza

Podstawowe umiejętności w zakresie przetwarzania i analizy danych w R (operacje na ramkach danych, obliczanie statystyk opisowych, estymacja modeli regresji). Podstawowa wiedza na temat wnioskowania statystycznego (estymacja średniej populacyjnej na podstawie prostej próby losowej).

## Wymagania wstępne do wykonania przed warsztatem

Instalacja pakietów survey i ggplot2.

# Analiza danych sondażowych w R

**Bartosz Sękiewicz**

HTA Consulting

**Opis warsztatu**

Celem warsztatu jest pokazanie z jakimi problemami możemy spotkać się podczas scrapowania stron www przy użyciu pakietu rvest. Warsztat pozwoli uczestnikom na uświadomienie sobie tego jak różnorodne mogą być strony internetowe (w kontekście ich konstrukcji). Dzięki poznaniu niuansów związanych z web scrapingiem możliwe będzie zaoszczędzenie w przyszłości sporej ilości czasu i nerwów. Z uwagi na ograniczoną ilość czasu pominiemy temat scrapowania stron obsługiwanych przez skrypty JS (wymaga to zastosowania dodatkowego oprogramowania jak PhantomJS, lub innego typu webscrapera jak RSelenium).

**Plan warsztatu** Podczas spotkania postaramy się rozwiązać problemy z pobieraniem danych ze stron zaproponowanych przez uczestników. Skupimy się na trzech aspektach:

1. piękno języka css, czyli wyciąganie informacji z kodu strony (m.in. tagi, klasy, id, rodzice i dzieci, sąsiedzi);
2. komunikacja ze stronami oraz nawigacja po nich (m.in. formularze, POST i GET);
3. API, czyli jak zaoszczędzić sobie czas (niestety nie zawsze jest to prawdziwe).

**Wymagane pakiety**

rvest (wystarczy zapoznanie się z opisem pakietu i jego zrozumienie, https://github.com/hadley/rvest)

**Wymagane od uczestników umiejętności i wiedza**

Podstawowa znajomość html, css. Mile widziana znajomość wyrażeń regularnych.

**Wymagania wstępne do wykonania przed warsztatem**

Przesłanie co najmniej trzech propozycji stron, którymi uczestnik byłby zainteresowany pod kątem web scrapingu. W zależności od przesłanych propozycji być może będzie konieczne założenie konta developerskiego dla wybranych serwisów (np. facebook, google).

# Web scraping w R i nie tylko

**Magdalena Mazurek**

Koło Naukowe Data Science

**Opis warsztatu**

Celem warsztatu jest zaprezentowanie możliwości pakietu RSelenium. Przedstawienie krótko jego wad oraz zalet. Uczestnicy z warsztatów dowiedzą się jak scrapować informacje ze stron internetowych wykorzystujących javascript oraz czemu warto przy tym używać zewnętrznej aplikacji PhantomJS.

**Plan warsztatu** Warsztaty rozpoczniemy od zaznajomienia uczestników z zasadą działania RSelenium oraz czym różni się od pakietu rvest. Zaczniemy od korzystania z RSelenium z użyciem klasycznej przeglądrki. W pierwszej kolejności zajmiemy się krótko scrapowaniem stron statycznych, niekorzystajacych z javascriptu jako prezentacja, że tradycyjne scrapowanie jest również możliwe, powiemy jednak czemu jest to nieefektywne. Następnie przejdziemy do części głównej, tj. scrapwowania stron korzystajacych z javascriptu, powiemy w tym miejscu czemu RSelenium jest możliwe do wykonywania tego. Na próbnej stronie pokażemy w jaki sposób korzystamy z pakietu. Na koniec powiemy o możliwości użycia aplikacji PhaontomJS.

**Wymagane pakiety**

RSelenium

**Wymagane od uczestników umiejętności i wiedza**

Podstawowa znajomość R i HTML.

**Wymagania wstępne do wykonania przed warsztatem**

Instalacja aplikacji PhantomJS, najnowszej wersji Java

# MicrosoftML - State of the art Machine Learning Microsoft

**Łukasz Grala**

Politechnika Poznańska Wydział Informatyki / TIDK - Data Scientist as a Services

### Opis warsztatu

Firma Microsoft kupiła firmę Revolution i od tego momentu oferuje produkt R Server. Najnowsza odsłona tego produktu R Server 9.0 dostępna na różne platformy SQL Server (Windows i Linux), Hadoop, Teradata, Spark zawiera między innymi nową bibliotekę MicrosoftML. Biblioteka ta jest podsumowaniem pracy naukowej Microsoft Research w zakresie Machine Learningu. Są tam między innymi wydajne algorytmy do klasyfikacji, szukania anomalii, czy regresji. Dostępne są tam również algorytmy Deep Learning wykorzystujące GPU.

### Plan warsztatu

1. Wprowadzenie do R Server
2. Algorytmy w MicrosoftML
3. Przykładowe scenariusze
4. Demonstracja Deep Learning z GPU

### Wymagane pakiety

MicrosoftML (R Server - może być zainstalowany trial, lub wersja developer z SQL Server - Linux lub Windows)

### Wymagane od uczestników umiejętności i wiedza

Podstawy języka R, znajomość podstawowych klas problemów i algorytmów uczenia maszynowego

### Wymagania wstępne do wykonania przed warsztatem

Instalacja R Server 9. W razie wybrania mojego warsztatu przygotuje do tego stosowny manual.

# Zastosowanie R w Power BI

**Dawid Detko**

Predica

### Opis warsztatu

Bardzo często potrzebujemy narzędzia, z którego ma korzystać ktoś co nie zna języków skryptowych, nie używa R Studio, czy notebooków. Możliwosć taką daje obecnie najbardziej popularny produkt na świecie to tzw Self-BI, czyli PowerBI firmy Microsoft. Produkt ten jest w pełni darmowy i daje możliwość zarówno pobierania danych ze skryptów w języku R, jak i tworzyć wizualizacje przy użyciu tego języka.

### Plan warsztatu

1. Przedstawienie PowerBI (w krótki i przystępny sposób)
2. Język R źródłem danych
3. Łączenie źródeł danych z języka R i innych źródeł
4. Wizualizacje w języku R

### Wymagane pakiety

Narzędzie PowerBI Desktop (darmowe), pakiety ggplot2, caret, lattice.

### Wymagane od uczestników umiejętności i wiedza

Podstawowa znajomość R

### Wymagania wstępne do wykonania przed warsztatem

PowerBI Desktop, R Studio

# Machine Learning w R przy użyciu H2O

**Michał Maj**

Appsilon Data Science, trigeR

### Opis warsztatu

Celem warsztatu jest zapoznanie uczestników z platformą H2O oraz dostępnymi algorytmami uczenia maszynowego jak np. Generalized linear model (GLM), Gradient Boosted Machines (GBM), Deep Neural Networks (DNN), K-means, Ensemble Methods.

### Plan warsztatu

1. Wprowadzenie - czym jest H2O i jak działa?
2. Przygotowanie i transformacje danych w H2O
3. Przegląd algorytmów + przykłady
4. Tuningowanie parametrów modelu
5. Ensemble Methods
6. Podsumowanie i dalsze wskazówki

### Wymagane pakiety

h2o, dplyr, ggplot2, h2oEnsemble

### Wymagane od uczestników umiejętności i wiedza

Podstawowa znajomość języka R. Mile widziana (choć niekonieczna) podstawowa wiedza z zakresu algorytmów uczenia maszynowego.

### Wymagania wstępne do wykonania przed warsztatem

R, RStudio, wymagane pakiety.

# Kombajn do uczenia maszynowego - MLR w praktyce

**Paweł Zawistowski**

Politechnika Warszawska, Wydział EiTI, AdForm

### Opis warsztatu

Celem warsztatu jest przedstawienie szerokich możliwości jakie daje pakiet MLR przy tworzeniu różnego rodzaju modeli - przejdziemy przy tym kompletną ścieżkę, od wstępnego przygotowania danych, przez wybór odpowiedniej metody, strojenie hiperparametrów, aż po diagnostykę i wizualizacje wyników.

### Plan warsztatu

1. Ogólne wprowadzenie do pakietu, przygotowanie środowiska MLR.
2. Przygotowanie danych do rozwiązywania naszego zadania (klasyfikacja, regresja, ...).
3. Wybór modelu, strojenie parametrów.
4. Diagnostyka i wizualizacja wyników.
5. Rozszerzanie MLR o własny algorytm.
6. Inne ciekawe elementy pakietu, podsumowanie.

### Wymagane pakiety

W ramach warsztatu korzystać będziemy z mlr oraz tidyverse. Udostępniony zostanie również obraz dockera ze wszystkim co potrzebne + RStudio.

### Wymagane od uczestników umiejętności i wiedza

1. Ogólna znajomość zagadnień związanych z tworzeniem modeli statystycznych, umiejętność korzystania z R'a w stylu "tidyverse".
2. Podstawowa umiejętność korzystania z dockera.

### Wymagania wstępne do wykonania przed warsztatem

Instalacja pakietów R lub ściągnięcie dockera i uruchomienie udostępnionego obrazu.

# XGBoost rządzi

**Vladimir Alekseichenko**

General Electric

## Opis warsztatu

XGBoost to jest jedna najlepszych implementacji "Gradient Boosting" z punktu widzenia praktycznego.9 Dlaczego warto?

1. Wynik (czyli zwykle jeden z najlepszych)
2. Czas na naukę i predykcję (potrafi używać wszystkie dostępne rdzenie)
3. Odporność na przeuczenia się (poprzez różne parametry regularyzacji)
4. Stabilność (można spokojnie używać na produkcje)

## Plan warsztatu

1. Zrozumienie biznes problemu
2. Zrozumienie danych
3. Budowa bardzo prostego modelu (base-line)
4. Przypomnienie co to jest drzewa decezyjne
5. Uruchomienie prostego modelu xgboost
6. Generowanie cech (feature engineering)
7. Budowanie bardziej zaawansowanego modelu
8. Optymalizacja hyperparametrów
9. Inne (zaawansowane) triki (opcjonalnie)

## Wymagane pakiety

xgboost, data.table, e1071, caret, rBayesianOptimization

## Wymagane od uczestników umiejętności i wiedza

Warsztat może być ciekawy dla osób które dopiero zaczynają, jak i dla średnio-zaawansowanych (z mojej wiedzy mało osób kojarzy i tym bardziej używa XGBoost w praktyce, chociaż to zmienia się bardzo szybko w czasie). Natomiast warto rozumieć podstawy:

1. uczenie maszynowe (machine learning)

2. cechy (features)
3. model, np. liniowy
4. przeuczenie się (overfitting)
5. walidacja (model evaluation)

Fajnie będzie jeżeli sprawdzisz (przypomnisz) jak działają drzewa decezyjne (decision trees).

### Wymagania wstępne do wykonania przed warsztatem

1. Mieć laptop z potrzebnymi pakietami (przede wszystkim xgboost)
2. Pobrać dane z Kaggle
3. Pomyśleć nad problemem przed warsztatem (może nawet spróbować go rozwiązać w najlepszy możliwy sposób - użyć dowolny model, który się zna)

# Klasyfikacja wieloetykietowa z pakietem R

**Paweł Teisseyre**

Instytut Podstaw Informatyki PAN

## Opis warsztatu

Celem warsztatów jest przedstawienie problemu klasyfikacji wieloetykietowej oraz pokazanie jak wykorzystać R do modelowania danych z wieloma etykietami. W klasycznym problemie klasyfikacji modelujemy zależność między zmienną odpowiedzi (najczęściej binarną) a zmiennymi objaśniającymi. W klasyfikacji wieloetykietowej rozważamy wiele binarnych zmiennych odpowiedzi jednocześnie. W ostatnich latach klasyfikacja wieloetykietowa wzbudziła bardzo duże zainteresowanie. Metody klasyfikacji wieloetykietowej są stosowane w wielu dziedzinach, takich jak automatyczna kategoryzacja tekstów, rozpoznawanie obrazów, modelowanie wielozachorowalności (współwystępowanie wielu chorób jednocześnie), przewidywanie skutków ubocznych leków i wiele innych. Podczas warsztatów opowiem o popularnych metodach stosowanych w klasyfikacji wieloetykietowej (takich jak łańcuchy klasyfikatorów). Podczas części praktycznej zajmiemy się analizą rzeczywistych zbiorów danych.

## Plan warsztatu

1. Teoria (omówienie problemu, przegląd metod).
2. Analiza danych rzeczywistych

## Wymagane pakiety

mldr

## Wymagane od uczestników umiejętności i wiedza

Podstawowa znajomość metod klasyfikacji i regresji.

## Wymagania wstępne do wykonania przed warsztatem

Brak

# Interaktywne wizualizacje w R i plotly - case study

**Piotr Ocalewicz**

Ocado Technology

## Opis warsztatu

Celem warsztatu jest zapoznanie użytkowników z możliwościami tworzenia interaktywnych wizualizacji danych korzystając z połączenia środowisk R, plotly oraz leaflet. To świetna okazja, żeby rozszerzyć swoje umiejętności w zakresie wizualizacji danych i nauczyć się tworzyć ciekawe i niebanalne podsumowania swojej pracy.

Warsztat prowadzony będzie w formie 'case study' - przejdziemy krok po kroku przez kolejne kroki analizy od krótkiego zapoznania się z danymi, poprzez stworzenie różnych interaktywnych wizualizacji aż po rozwiązanie problemu, który przed sobą postawiliśmy.

W trakcie warsztatu stworzymy kompletny dokument w formacie html zawierający podsumowanie analizowanych danych oraz stworzone przez nas grafiki. Omówimy zarówno podstawowe rodzaje wykresów, te bardziej zaawansowane jak również sposoby ich połączenia w jednym, estetycznym podsumowaniu.

W trakcie szkolenia każdy uczestnik otrzyma wydrukowane 'ściągawki' zawierające najważniejsze funkcje i składnię omawianych pakietów.

## Plan warsztatu

1. Omówienie zbioru danych i problemu do rozwiązania
2. Krótkie wprowadzenie do pakietów ggplot2 oraz rmarkdown
3. Środowisko plotly i jego współpraca z R
4. Podstawowe typy wykresów
5. Zaawansowane wykresy i sposoby ich edycji
6. Dynamiczne zmienianie zawartości wykresów - guziki, suwaki itd.
7. Interaktywna wizualizacja na mapach
8. Podsumowanie warsztatu i wyników analizy

## Wymagane pakiety

dplyr, ggplot2, rmarkdown, knitr, plotly, ggmap, leaflet, flexdashboard, ggiraph

**Wymagane od uczestników umiejętności i wiedza**

Umiejętność tworzenia wykresów w R i co najmniej podstawowa znajomość pakietów ggplot2 i dplyr. Mile widziana znajomość markdown.

**Wymagania wstępne do wykonania przed warsztatem**

Pakiety do zainstalowania: dplyr, ggplot2, rmarkdown, knitr, plotly, ggmap, leaflet, flexdashboard, ggiraph. Zainstalowane środowiska RStudio. Darmowe konto w serwisie www.plot.ly

# Efektywna i efektowna wizualizacja w ggplot2

**Piotr Ćwiakowski**

Uniwersytet Warszawski

### Opis warsztatu

Przedstawienie zaawansowanych funkcji i rozszerzeń pakietu ggplot2. Po warsztacie użytkownik zna zaawansowane możliwości pakietu ggplot2 (m. in. interaktywne wykresy) oraz poznał zasady poprawnej wizualizacji danych

### Plan warsztatu

1. Wprowadzenie do tidyverse, grammar of graphics i ggplot2
2. Zasady działania ggplot2
3. Przegląd geometrii w ggplot2 (z szczególnym uwzględnieniem zaawansowanych i nietypowych
4. Przegląd rozszerzeń do ggplot2
5. Sztuka tworzenia wykresów
6. Tworzenie złożonych i zaawansowanych wykresów w ggplot2 w praktyce

### Wymagane pakiety

tidyverse, ggplot2 extensions

### Wymagane od uczestników umiejętności i wiedza

Podstawowy R, mile widziane doświadczenie w analizie danych (niekoniecznie w R)

### Wymagania wstępne do wykonania przed warsztatem

Zainstalowanie R (z opcjonalną, ale rekomendowaną nakładką R Studio), zainstalowanie pakietu tidyverse i wybranych pakietów z rodziny ggplot2 extensions

# Wróżenie z punktów - ordynacja w eksploracji danych

**Marcin K. Dyderski**

Instytut Dendrologii Polskiej Akademii Nauk

## Opis warsztatu

Warsztaty zakładają wprowadzenie do technik ordynacji - uporządkowania punktów w przestrzeni cech i redukcji wielowymiarowości do postaci zdatnej do wyrażenia za pomocą prostego wykresu. Ordynacja może być sama w sobie metodą do wykazania pewnych prawidłowości, może też jednak stanowić źródło do wyszukiwania zależności, które chcemy/musimy przedstawić później w bardziej wysublimowany sposób.

Celem warsztatów jest wskazanie możliwości zastosowania podstawowych technik ordynacyjnych do wyszukania zależności pomiędzy danymi. Planuję podczas warsztatów przeanalizować wraz z Uczestnikami trzy zbiory danych, w których będziemy szukać zależności związanych z problemem analitycznym. Oprócz tego chciałbym krótko omówić przykłady zastosowania tego typu analiz oraz najczęściej popełniane błędy.

Uczestnicy podczas warsztatów nauczą się:

1. jak przygotować dane do analiz z zastosowaniem ordynacji
2. w jaki sposób wykonać analizę głównych składowych (PCA), analizę korespondecji (CA) oraz kanoniczną analizę korespondencji (CCA)
3. w jakich warunkach dana analiza może być zastosowana, jakie ma wady oraz jakie ma ograniczenia
4. w jaki sposób interpretować uzyskane wyniki oraz jak przedstawić je graficznie w sposób przystępny i estetyczny

## Plan warsztatu

1. Czym jest ordynacja - wprowadzenie
2. Podział metod, zastosowania i przykłady
3. Przygotowanie danych, transformacje
4. Przypadek 1 - czym się różnią drzewa?
5. Przypadek 2 - co wpływa na naszą ocenę piwa?
6. Przypadek 3 - czym się różnią od siebie miasta?
7. Podsumowanie + informacje gdzie szukać dalej

**Wymagane pakiety**

vegan, ggplot2, gridExtra, scales

**Wymagane od uczestników umiejętności i wiedza**

Podstawowa znajomość R: operacje na data.frame'ach, macierzach i listach, umiejętność tworzenia wykresów w pakiecie ggplot2

**Wymagania wstępne do wykonania przed warsztatem**

komputer z R oraz zainstalowanymi pakietami

# mirt: skalowanie odpowiedzi lepsze niż PCA

**Piotr Migdał**

deepsense.io, freelancer

## Opis warsztatu

Item Response Theory jest modelem analizy danych, w której szukamy zmiennej ukrytej wyjaśniającej dane. Np. zamieniamy wiele odpowiedzi z ankiety na jedną zmienną odpowiadającą zadowoleniu klienta, czy też estymujemy pewną cechę charakteru na podstawie kwestionariusza. Innym zastosowaniem jest skalowanie wyników egzaminów w sposób mądrzejszy niż liczenie sumy punktów (nie każde zadanie jest równie trudne, niektóre zadania mogą być losowe).

Typowe sposoby (np. liczenie pierwszej składowej w PCA) nie uwzględniają nieliniowości zmiennych.

Pakiet mirt (Multidimensional Item Response Theory) jest wydajnym i wszechstronnym pakietem do praktycznych zastosowań IRT.

## Plan warsztatu

1. wprowadzenie do Item Response Theory
2. różne modele zmiennych odpowiedzi (też: gradualne)
3. szukanie zmiennej ukrytej
4. generowanie sztucznych odpowiedzi
5. ćwiczenie praktyczne: analiza danych maturalnych

## Wymagane pakiety

mirt, ggplot2, dplyr

## Wymagane od uczestników umiejętności i wiedza

Podstawy R. Co to jest sigmoida.

## Wymagania wstępne do wykonania przed warsztatem

mirt; RStudio z R Notebook (lub ew. R w Jupyter Notebook)

# Social Network Analysis w R

**Michał Wojtasiewicz**

Instytut Podstaw Informatyki PAN

## Opis warsztatu

Celem warsztatu jest zaznajomienie uczestników z tematyką analizy danych w sieciach społecznych. Ćwiczenia praktyczne zostaną przeprowadzone w głównej mierze przy użyciu pakietu igraph. Dziedzina Social Network Analysis (SNA) jest teraz jedną z najprężniej rozwijających się dziedzin uczenia maszynowego. Z racji powszechności występowania sieci społecznych (np. Facebook, Instagram, LinkedIn, problemy optymalizacyjne, sieci systemów rekomendujących, sieć połączeń mailowych) zapotrzebowanie na algorytmy i coraz bardziej zaawansowane rozwiązania stale wzrasta. Naturalną metodą zapisu sieci jest graf czyli zbiór wierzchołków i krawędzi. Dzięki łatwej w konstrukcji strukturze, zapis grafowy pozwala na skuteczne rozwiązywanie szerokiego zakresu problemów data miningowych. Na zajęciach warsztatowych uczestnicy zapoznają się z tematyką SNA, głównym problemami oraz popularnymi rozwiązaniami tych problemów. Nauczą się wyznaczać grupy podobnych elementów sieci (np. grupa znajomych), kluczowe ze względu przesyłania informacji elementy sieci (np. bottlenecki), podgrupy elementów pozornie niepowiązanych (np. grupa klientów kupujących ten sam produkt) oraz użycia sieci do szeregowania zadań (kolorowanie zwarte).

## Plan warsztatu

1. Wprowadzenie do tematyki SNA.
2. Omówienie przykładowej sieci poprzez strukturę grafu.
3. Analiza skupień w sieci społecznej.
4. Wyznaczenie różnych rodzajów najbardziej wpływowych elementów sieci.
5. Wprowadzenie do systemów rekomendujących.
6. Szeregowania zadań z restrykacją braku przestoju.

## Wymagane pakiety

igraph", Matrix, visNetwork

## Wymagane od uczestników umiejętności i wiedza

Podstawowa znajomość języka R.

## Wymagania wstępne do wykonania przed warsztatem

Pobranie przykładowego grafu oraz wymaganych pakietów.

# Text mining w R

**Norbert Ryciak**

Politechnika Warszawska, Wydział MiNI

## Opis warsztatu

Celem warsztatu jest zapoznanie uczestników z podstawowymi technikami stosowanymi podczas analizy tekstu. Omówione zostanie m. in. zagadnienie modelowania tematycznego przy użyciu modelu LDA. Duży nacisk będzie położony na poznanie specyfiki pracy z danymi tekstowymi i zrozumienie motywacji prowadzących do określonych metod analizy. Wybrane zagadnienia zostaną zaprezentowane w zastosowaniu do grupowania lub klasyfikacji tekstów.

## Plan warsztatu

1. Wstępne przetwarzanie i redukcja wymiaru danych
2. Podstawowe metody reprezentacji zbioru danych tekstowych
3. Modelowanie tematyczne - model LDA (Latent Dirichlet Allocation)

## Wymagane pakiety

tm, topicmodels, twitteR, graph, Rgraphviz, ggplot2, LDAvis

## Wymagane od uczestników umiejętności i wiedza

1. Podstawowa umiejętność programowania w R
2. Wiedza czym jest klasyfikacja statystyczna i analiza skupień

## Wymagania wstępne do wykonania przed warsztatem

brak

# Kiedy brakuje wydajności... R i C++ = Rcpp

**Zygmunt Zawadzki**

zstat

## Opis warsztatu

Celem warsztatu jest nauczenie użytkowników wykorzystania pakietu Rcpp pozwalającego wykorzystać kod C++ w R w celu przyspieszenia krytycznych fragmentów obliczeń.

Uczestnik po skończonym warsztacie będzie potrafił:

1. przedstawić różnice w modelu zarządzania pamięcią w R i C++ i omówić konsekwencje które się z tym wiążą.
2. stworzyć prosty pakiet R wykorzystujący kod C++.
3. wykorzystać pakiet profvis do wyszukania najbardziej gorącego fragmentu kodu, który potencjalnie mógłby zostać przepisany z wykorzystaniem Rcpp.

## Plan warsztatu

Wprowadzenie do Rcpp: Część I:

1. Kompilacja pierwszej funkcji wykorzystującej C++ w R.
2. Omówienie różnic pomiędzy językiem interpretowanym i kompilowanym na przykładzie R i Rcpp.
3. Szczegółowe omówienie struktur R dostępnych w C++ (NumericVector i NumericMatrix)
4. Przedstawienie STL - standardowej biblioteki szablonów jako dodatkowych struktur danych gotowych do wykorzystania.
5. Praktyczne prezentacja modeli zarządzania pamięcią w C++ - stworzenie kilku mini-funkcji prezentujących możliwe konsekwencje błędnej interakcji R i C++.

Część II:

1. Profilowanie kodu z wykorzystaniem Rprof.
2. Wprowadzenie biblioteki RcppArmadillo do obliczeń macierzowych w C++.
3. Stworzenie prostego samplera Gibbsa wykorzystującego RcppArmadillo i funkcje R dostępne po stronie C++.

## Wymagane pakiety

Rcpp, RcppArmadillo, profvis

## Wymagane od uczestników umiejętności i wiedza

Podstawy programowania w R:

1. operacje na macierzach.
2. pętle for.

Znajomość C++ nie jest wymagana. Wszystkie potrzebne informacje dotyczące tego języka zostaną omówione w trakcie warsztatów.

## Wymagania wstępne do wykonania przed warsztatem

W przypadku systemu Windows pobranie i instalacja: Rtools - najnowsza wersja (https://cran.r-project.org/bin/windows/Rtools/)

Linux: wszystko powinno być zainstalowane (potrzebny jest kompilator gcc z obsługą standardu C++11 - jednak wszystkie w miarę nowe wersje powinny go mieć).

Mac: zainstalowane XCode.

# Nie pisz kodu, pisz prozę - wprowadzenie do pakietu dplyr

**Bartłomiej Tartanus**

OSA/Sages

**Opis warsztatu**

Ramki danych w R (data.frame) są niezbędne do pracy z danymi w postaci tabelarycznej. Jednak często przetwarzanie takich ramek przy użyciu czystego R prowadzi do wielokrotnego powtarzania nazw ramki czy kolum w celu np przefiltrowania wierszy lub niewygodnego doklejania kolumn w przypadku rozszerzania ramki. Taki kod często bywa nieczytelny i nie do końca oddaje intencje autora. Do takiego kodu ciężko wrócić za jakiś czas i przypomnieć sobie "co ja tutaj chciałem zrobić". Wtedy mamy faktycznie do czynienia z "kodem". Czy tak musi być już zawsze? Na szczęście nie. Z pomocą przychodzi pakiet dplyr, który pozwoli nam pisać "prozę" - nasz kod będzie o wiele czytelniejszy.

**Wymagane pakiety**

dplyr

**Wymagane od uczestników umiejętności i wiedza**

Podstawowa znajomość R - operacje na wektorach, tworzenie ich oraz znajomość ramek danych.

**Wymagania wstępne do wykonania przed warsztatem**

R, RStudio

# Indeks nazwisk