

Ogólnopolska Konferencja Użytkowników R



Warszawa, 27-29 września 2017



**Wydział Matematyki
i Nauk Informacyjnych**

POLITECHNIKA WARSZAWSKA



Why R? 2017

Ogólnopolska Konferencja Użytkowników R

Warszawa, 27-29 września 2017

Spis treści

1 Wykłady plenarne	19
Dlaczego czasem R? Why (sometimes) R?	20
BioinfoRmatyka nowotworów	21
R jako główna platforma do zaawansowanej analityki w Enterprise	22
Jak dużo mocy (i po co) można wycisnąć z modelu predykcyjnego?	23
Sorted L-One Penalized Estimation	24
Dezagregacja danych przedziałowych	25
Kim naprawdę był Gall Anonim? Zagadnienia statystycznej analizy tekstu	26
Metody wizualizacji danych jakościowych w programie R	27
Analiza danych obserwacyjnych zwiedzających wystawy w Centrum Nauki Kopernika	28
Show Me Your Model	29
2 Sesje	31
PwC Data Analytics	32
Tidyverse	34
Modelowanie Ryzyka (UBS)	36
Statystyka	38
Biostatystyka i Edukacja	42
Biznes	44
Społeczności	46
Panel dyskusyjny: Edukacja statystyki	46
3 Lightning Talks	47
4 Sesja plakatowa	51
5 Warsztaty	55
Analiza danych sondażowych w R	56
Złożone schematy doboru próby - pakiet survey	57
Pakiet rvest, czyli web scrapingu wybrane przypadki	58
Web scraping w R i nie tylko	59
MicrosoftML - State of the art Machine Learning Microsoft	60
Zastosowanie R w Power BI	61
Machine Learning w R przy użyciu H2O	62
Kombajn do uczenia maszynowego - MLR w praktyce	63
XGBoost rządzi	64
Klasyfikacja wieloklasowa z pakietem R	65
Interaktywne wizualizacje w R i plotly - case study	66
Efektywna i efektowna wizualizacja w ggplot2	67
Wróżeńie z punktów - ordynacja w eksploracji danych	68
mirt: skalowanie odpowiedzi lepsze niż PCA	69
Social Network Analysis w R	70
Text mining w R	71
Kiedy brakuje wydajności... R i C++ = Rcpp	72
Nie pisz kodu, pisz prozę - wprowadzenie do pakietu dplyr	73
Indeks nazwisk	75

Komitet organizacyjny



Marcin Kosiński
(Przewodniczący)
Gradient Metrics



Olga Mierzwa-Sulima
Appsilon



Przemysław Biecek
*Uniwersytet Warszawski,
Politechnika Warszawska,
MI²*



Bartosz Sękiewicz
HTA Consulting



Maciej Beręsewicz
*Uniwersytet Ekonomiczny
w Poznaniu*



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



Alicja Gosiewska
*Politechnika Warszawska,
MI²*



Aleksandra Dąbrowska
*Uniwersytet Warszawski,
MI²*



Monika Stępień
Uniwersytet Warszawski



Agnieszka Dumania
SalesBI



Anna Rybińska
Uniwersytet Gdańskiego



Bartłomiej Tartanus
OSA/Sages



Krzysztof Słomczyński
BD Polska



Emil Buszyło
Maxus Net Communication



Konrad Więcko
Samsung

Skład książki: Alicja Gosiewska, Aleksandra Dąbrowska

Korekta: Marcin Kosiński

Okładki: Klaudia Korniluk (facebook.com/pedzlenie)

Wstęp

Nie ukrywam ogromnych nadziei jakie wiążę z tym wydarzeniem. Jest to wyjątkowa szansa na wzmacnienie całej polskiej społeczności analityków danych i pasjonatów języka R. Takie spotkania są potrzebne nam wszystkim, na co dzień zmagającym się z wyzwaniami stawianymi przez świat analizy danych.

To tu praktycy ze środowiska biznesowego mogą pochwalić się osiągnięciami i zasięgnąć rady od kolegów z branży. To tu naukowcy i przedstawiciele świata akademickiego mają możliwość przedstawienia wyników badań oraz projektów, nad którymi pracują. Ostatecznie, to tu wszyscy rozumiemy, że tylko poprzez wspólną wymianę doświadczeń i przemyśleń jesteśmy w stanie razem utrzymywać renomę polskich specjalistów w dziedzinie analizy danych na światowym poziomie.

Marcin Kosiński

Przewodniczący Komitetu Organizacyjnego
Ogólnopolskiej Konferencji Użytkowników R
Why R? 2017

Współorganizator
Warszawskich Spotkań Entuzjastów R

Pasjonat R, analizy danych
i organizacji wydarzeń

Plan konferencji

Dzień warsztatowy 27.09	9:00 – 10:30	Sesja poranna	
	10:30 – 11:00	Przerwa kawowa	
	11:00 - 12:30	Sesja poranna – ciąg dalszy	
	12:30 – 14:00	Przerwa na lunch	
	14:00 – 15:30	Sesja popołudniowa	
	15:30 – 16:00	Przerwa kawowa	
	16:00 – 17:30	Sesja popołudniowa – ciąg dalszy	
Dzień 1 28.09	8:00 – 9:00	Rejestracja	
	9:00 – 9:30	Inauguracja (107)	
	9:30 – 10:30	Keynotes: Tomasz Burzykowski & Ewa Szczurek (107)	
	10:30 – 11:00	Przerwa kawowa	
	11:00 – 12:30	Tidyverse (102)	PwC Data Analytics (107)
	12:30 – 13:30	Keynotes: Wit Jakuczun & Artur Suchwałko (107)	
	13:30 – 14:30	Lunch (PwC)	
	14:30 – 16:00	Statystyka (102)	Modelowanie ryzyka (UBS) (107)
	16:00 – 16:30	Przerwa kawowa (Pearson IOKI)	
	16:30 – 17:30	Keynotes: Małgorzata Bogdan & Michał Ramsza (107)	
	17:30 – 18:30	Panel dyskusyjny: Edukacja statystyki (107)	
	18:30 – 20:00		
	20:00 – ∞	Welcome paRty (Aula Gmachu Fizyki) Sesja plakatowa	
Dzień 2 29.09	8:00 – 9:00	Rejestracja	
	9:00 – 10:00	Keynotes: Maciej Eder & Justyna Brzezińska (107)	
	10:00 – 11:30	Biostatystyka i Edukacja (102)	Biznes (107)
	11:30 – 12:00	Przerwa kawowa	
	12:00 – 13:00	Keynotes: Anna Wróblewska & Przemysław Biecek (107)	
	13:00 – 14:00	Lunch (UBS)	
	14:00 – 15:00	Sesja sponsorska (107)	
	15:00 – 15:30	Przerwa kawowa (Sages)	
	15:30 – 16:00	Lightning Talks (107)	
	16:00 – 17:00	Sesja społeczności (107)	
	17:00 – 17:30	Zakończenie (107)	

Plan warsztatów

Sesja	Sala	Sesja poranna 9:00 – 12:30	Sesja popołudniowa 14:00 – 17:30
Sondaże	211	Analiza danych sondażowych w R Dariusz Szklarczyk Agnieszka Otręba - Szklarczyk, Fundacja Rozwoju Badań Społecznych	Złożone schematy doboru próby - pakiet survey Tomasz Żółtak, Instytut Badań Edukacyjnych
Web-harvesting	212	Pakiet rvest, czyli web scrapingu wybrane przypadki Bartosz Sękiewicz, HTA Consulting	Web scraping w R i nie tylko Magdalena Mazurek, Koło Naukowe Data Science
Uczenie Maszynowe, Microsoft / Self-Service BI, Microsoft	213	MicrosoftML - State of the art Machine Learning Microsoft Łukasz Grala, Politechnika Poznańska Wydział Informatyki / TIDK - Data Scientist as a Services	Zastosowanie R w Power BI Dawid Detko, Predica
Uczenie Maszynowe, Narzędzia	214	Machine Learning w R przy użyciu H2O Michał Maj, Appsiilon Data Science, trigeR	Kombajn do uczenia maszynowego - MLR w praktyce Paweł Zawistowski, Politechnika Warszawska, Wydział EiT / AdForm
Uczenie Maszynowe, Metody	311	XGBoost rządzi Vladimir Alekseichenko, GE Healthcare	Klasyfikacja wieloetykietowa z pakietem R Paweł Teisseyre, Instytut Podstaw Informatyki PAN
Wizualizacje	312	Interaktywne wizualizacje w R i plotly - case study Piotr Ocalewicz, Ocado Technology	Efektywna i efektowna wizualizacja w ggplot2 Piotr Ćwiakowski, Uniwersytet Warszawski
Analiza wielowymiarowa	313	Wróżenie z punktów - ordynacja w eksploracji danych Marcin K. Dyderski, Instytut Dendrologii Polskiej Akademii Nauk	mirt: skalowanie odpowiedzi lepsze niż PCA Piotr Migdał, deepsense.io / freelancer
Grafy / Analiza Tekstu	314	Social Network Analysis w R Michał Wojtasiewicz, Instytut Podstaw Informatyki PAN	Text mining w R Norbert Ryciąk, Politechnika Warszawska, Wydział MiNI
Wydajność	316	Kiedy brakuje wydajności... R i C++ = Rcpp Zygmunt Zawadzki, zstat	Nie pisz kodu, pisz prozę - wprowadzenie do pakietu dplyr Bartłomiej Tartanus, OSA/Sages

Organizatorzy



**Wydział Matematyki
i Nauk Informacyjnych**

POLITECHNIKA WARSZAWSKA

Wydział Matematyki i Nauk Informacyjnych Politechnika Warszawska

Wydział Matematyki i Nauk Informacyjnych powstał w 1999 roku w wyniku podziału Wydziału Fizyki Technicznej i Matematyki Stosowanej. Początkowo miał siedzibę w Gmachu Głównym Politechniki Warszawskiej, a w 2012 roku przeniósł się do nowo wybudowanego budynku przy ul. Koszykowej 75.

Wydział może pochwalić się wysoką pozycją naukową (kategoria A) oraz wysokim poziomem kształcenia (wyróżnienie na kierunku Matematyka i pozytywna ocena na kierunku Informatyka Państwowej Komisji Akredytacyjnej), zaangażowaniem kadry naukowo-dydaktycznej w działalność publikacyjną oraz prowadzeniem dużych projektów informatycznych, w tym projektów zakończonych wdrożeniami.

Wydział prowadzi studia pierwszego i drugiego stopnia na kierunkach Matematyka, Informatyka i Computer science (studia w języku angielskim) oraz studia niestacjonarne pierwszego stopnia na kierunku Matematyka, a także studia doktoranckie z matematyki. Wydział uruchamia studia doktoranckie z informatyki od października 2015 roku. Obecnie na Wydziale studiuje około 1000 studentów.

Wydział MiNI posiada prawo do nadawania stopni doktora i doktora habilitowanego w dziedzinie matematyki oraz stopnia doktora w dziedzinie informatyka. W wyniku podjętych działań nastąpiła znacząca poprawa bazy lokalowej Wydziału (nowy gmach). Umożliwiło to rozwój bazy laboratoryjnej Wydziału, zwiększenie naboru na kierunku Informatyka i Computer Science oraz stabilizację kadry naukowo-dydaktycznej.

Absolwenci kierunku **Matematyka** posiadają wiedzę niezbędną do formułowania oraz rozwiązywania problemów teoretycznych, tworzenia i rozwiązywania matematycznych modeli zjawisk w różnych dziedzinach życia. Są przygotowani do pracy w bankach i innych instytucjach sektora finansowego, urzędach statystycznych, administracji państowej oraz placówkach naukowych. Absolwenci kierunku **Informatyka**, poza wiedzą informatyczną, posiadają gruntowne przygotowanie matematyczne. Znajdują zatrudnienie jako kierownicy zespołów programistycznych, projektanci oprogramowania i sieci komputerowych, administratorzy systemów informatycznych, czy jako specjaliści ds. ochrony danych i bezpieczeństwa informacji.



Grupa MI²

MI² (czytaj: mi kwadrat) powstała jako pomost pomiędzy MIM UW i MiNI PW, czyli dwoma wydziałami pełnymi pasjonatów analizy danych oraz tworzenia narzędzi matematycznych i informatycznych do analizy danych. Zaczęło się od UW i PW ale w grupie działały też osoby z innych uczelni, miast czy nawet krajów zarówno studenci jak i absolwenci.

Działamy w modelu think & do. Analiza danych nie jest wartością samą w sobie. Ważne by była odpowiedzią na rzeczywiste potrzeby i prowadziła do lepszych decyzji. Rozwijamy umiejętności identyfikacji problemu, rozwiązywania problemu i wdrażania rozwiązania w oparciu o dane i zaawansowane metody analityczne, skalowane narzędzia informatyczne. Robimy to realizując projekty przy których nie tylko można się czegoś dowiedzieć, ale też można coś użytecznego zrobić a wyniki analiz przełożyć na faktyczne akcje.



Koło Naukowe Data Science

Koło Naukowe Data Science działa przy Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej, który wypuszcza w świat nie tylko świętych matematyków i informatyków, ale także Data Scientistów! Patrz nasze specjalności:

- na studiach magisterskich z matematyki - ciesząca się prestiżem wśród pracowników **Statystyka Matematyczna i Analiza Danych**,
- na studiach magisterskich z informatyki - właśnie uruchomiona specjalność **Przetwarzanie i Analiza Danych**.

Nasze Koło zrzesza osoby zainteresowane szeroko pojętą analizą danych. Tworzymy grupę ambitnych studentów, którzy rozwijają się przez podejmowanie działań wykraczających ponad to, co dostarczają nam studia. Uczestnictwo w życiu Koła daje nam możliwości nabycia konkretnych umiejętności, poznawania przydatnych narzędzi, zdobywania cennego doświadczenia i doskonalenia umiejętności miękkich. Jednym z głównych elementów naszej działalności jest organizacja warsztatów.

Partnerzy społecznościowi



Szychta w danych

Szychta w danych - ważne tematy, otwarte dane, udostępnione kody, niebanalne wnioski. Każdy wpis Szuchty to historia o świecie, zbudowana wokół danych. Narracja zbudowana jest prostym językiem - bez zbędnych ozdobników, bez okładania się ideologiczną maczugą, bez półprawd i bez post-prawdy. Świat czeka na to, aby go poznać!



Tableau & Data Viz Meetup

Tableau & Data Viz Meetup jest grupą użytkowników Tableau i entuzjastów wizualizacji danych. Celem spotkań jest popularyzacja dobrych praktyk tworzenia wykresów i raportów oraz umiejętności analitycznych w programie Tableau.



QuantFin



Stowarzyszenie Wrocławskich Użytkowników R (STWUR)

Stowarzyszenie Wrocławskich Użytkowników R to dolnośląska grupa miłośników R. Na spotkaniach stawiamy sobie trzy cele. Pierwszy to integracja eRowego środowiska, ludzi, którzy w R programują lub są nim zainteresowani. Drugi to wspólna analiza inspirujących zbiorów danych, wyciąganie wniosków i komunikowanie ich poza grupę. Last but not least, spotkania STWURa to idealne miejsce na wymianę doświadczeń, naukę nowych pakietów i szansę na (lepsze) poznanie R.

W przeciwieństwie do większości spotkań użytkowników R, STWUR jest zaplanowany jako cykl spotkań, których wspólnym motywem jest jeden zbiór danych. Zaczynamy od danych z Diagnozy Społecznej, które mamy zamiar ze spotkania na spotkanie coraz głębiej eksplorować i poznawać.

>eRka()

Entuzjastów R
Krakowska Alternatywa

entuzjastów R krakowska alternatywa (eRka)

eRka to cykliczne spotkania osób zainteresowanych obszarami analizy danych, data science i Big Data.

Chcemy zaoferować analitykom danych (oraz osobom, które chcą nimi zostać) przestrzeń, pozwalającą m.in. na zdobycie ciekawych kontaktów, wiedzy, czy też cennego doświadczenia. Wszystkie nasze działania w mniejszym bądź większym stopniu powiązane są także z promocją języka R – obecnie najszybciej rozwijającego się środowiska do obliczeń statystycznych.

trige(R)

Trójmiejska Grupa Entuzjastów R (trigeR)

trigeR, czyli Trójmiejska Grupa Entuzjastów R to nowo powstała, dynamicznie rozwijająca się grupa działająca na Pomorzu i zrzeszająca osoby, które na co dzień zajmują się szeroko pojętą analizą danych - zarówno w branży biznesowej, jak i na polu akademickim. trigeR został założony w grudniu 2016 roku przez Annę Rybińską, doktorantkę Pracowni Chemometrii Środowiska Uniwersytetu Gdańskiego oraz Agnieszkę Borsuk i Emilię Daghir-Wojtkowiak, doktorantki Katedry Biofarmacji i Farmakokinetyki Wydziału Farmaceutycznego Gdańskiego Uniwersytetu Medycznego.



R-Ladies Warsaw

R-Ladies to ogólnoświatowa inicjatywa, promująca różnorodność płci w środowisku użytkowników R. R-Ladies Warsaw powstała przy Warszawskich Spotkaniach Entuzjastów R, z Olgą Mierzwą-Sulimą jako inicjatorką, stawiając sobie za cel rozpropagowanie środowiska R wśród kobiet.



Warszawskie Spotkania Entuzjastów R (SER)

Warszawskie Spotkania Entuzjastów R mają na celu integrację środowiska użytkowników i entuzjastów analizy danych z użyciem programu R. Spotkania organizowane są co miesiąc w czwartki w godzinach popołudniowych. Każde spotkanie składa się z dwóch półgodzinnych prelekcji rozdzielonych półgodzinną przerwą na spokojną dyskusję w kularach. Zapraszamy nowicjuszy i weteranów, każdy znajdzie coś dla siebie!



Poznański Akademicki Zlot Użytkowników R (PAZUR)

PAZUR (Poznański Akademicki Zlot Użytkowników R) to cyklicznie odbywające się spotkania zrzeszające pasjonatów języka R zarówno z biznesu, jak i środowiska akademickiego. Historia spotkań z cyklu PAZUR sięga roku 2012 i od tego czasu odbyło się 20 “małych” spotkań oraz dwie “duże” konferencje - Polski Akademicki Zlot Użytkowników R w 2014 roku oraz European R Users Meeting w 2016 roku. Organizowane prelekcje cieszą się dużym zainteresowaniem i stanowią doskonałą okazję do wymiany doświadczeń oraz nawiązania nowych kontaktów w świecie analizy danych. Spotkania PAZUR mogą się pochwalić wsparciem R Consortium na poziomie Matrix!



Fundacja Rozwoju Badań Społecznych (FuRBS)

Fundacja Rozwoju Badań Społecznych (FuRBS) to organizacja skupiona na doskonaleniu i promocji metod badań społecznych i analizy danych w życiu społeczno-gospodarczym. Dążymy do tego, by badania społeczne i analiza danych nadawały za postępem technologicznym i zmieniającymi się potrzebami społecznymi, a przy tym były dostępne i zrozumiałe dla szerokiego grona odbiorców. Prowadzimy prace badawczo-rozwojowe służące poszerzeniu stosowania badań społecznych i analizy danych w praktyce życia społecznego. Dziedziną szczególnego zainteresowania Fundacji są działania na rzecz wzrostu rangi i wykorzystywania w praktyce badań danych zastanych i metod niereaktywnych. Świadczymy również usługi konsultingu badawczego i analitycznego w dziedzinie badań naukowych, badań rynku i ewaluacji oraz usługi szkoleniowe z zakresu metod i technik badawczych, statystycznej analizy danych i obsługи narzędzi analitycznych.



Stacja IT

Stacja IT to inicjatywa organizowana siłami społeczności - miejsce spotkań ekspertów branżowych, profesjonalistów, którzy nieprzerwanie podnoszą swoje kompetencje, trenerów, przedsiębiorców i twórców startupów, a także osób dopiero wchodzących na rynek pracy. Dzięki unikalnemu składowi osobowemu oraz doświadczeniu partnerów, Stacja zbliża ze sobą świat biznesu, ośrodki naukowe i centra kompetencyjne, aby przekuwać wiedzę w realną wartość dla uczestników jej przedsięwzięć.



Laboratorium Cyfrowe Humanistyki Uniwersytetu Warszawskiego (LaCh UW)

To struktura w ramach Uniwersytetu Warszawskiego, wspierająca humanistów i humanistki realizujących cyfrowe projekty naukowe. Organizujemy także liczne warsztaty z podstawowych dla humanistyki cyfrowych metod i narzędzi. Jak dotąd dotyczyły one przetwarzania danych i tekstów, web scrappingu, tworzenia wydań cyfrowych czy organizacji i budowania nowoczesnych bibliotek oraz repozytoriów. W tym semestrze prowadzimy trzy cykle warsztatowe z podstaw R oraz grupę samokształceniową “Python w badaniach humanistycznych”.

Interesują nas nie tylko metody, ale szersze spojrzenie na kulturę cyfrową. W ramach cyklu “Poza interfejsem” organizujemy wykłady otwarte prowadzone przez osoby, które analizują technologie i oprogramowanie z humanistycznej perspektywy. Podczas tych spotkań sprawdzamy, jak algorytmy wpływają na nasze myślenie o świecie i co kryje się za nieprzejrzystymi interfejsami. Naszymi gośćmi byli już między innymi Aleksandra Przegalińska (ALK, MIT) - badaczka zajmująca się teoriami sztucznej inteligencji oraz Paweł Janicki (Centrum Sztuki WRO) - kurator i artysta, twórca interaktywnych instalacji i systemów.



Data Science Po Polsku

Data Science po polsku podcast przeznaczony dla wszystkich zainteresowanych Data Science, analizą danych i analizą biznesową. W krótkich, cotygodniowych odcinkach opowiadam o różnych tematach powiązanych z Data Science. Szczególną uwagę zwracam na wszystko co dotyczy polskiego rynku Data Science. Wierzę że Polska ma szanse (o ile już to nie nastąpiło) stać się światowym centrum Data Science.



SmarterPoland.pl

Smarter Poland



Data Science Warsaw



Microsoft Azure
User Group Poland

Microsoft Azure User Group Poland



Polish SQL Server User Group (PLSSUG)

datahero.tech

DataHero.Tech



Data Science Łódź

Wykłady plenarne

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.

Bio

Tomasz Burzykowski uzyskał dyplom magistra w zastosowaniach matematyki (1990) na Uniwersytecie Warszawskim, a także dyplom magistra (1991) i doktora (2001) w biostatystyce na Uniwersytecie Hasselt (Belgia).

Pracuje jako profesor biostatystyki/bioinformatyki na Uniwersytecie Hasselt oraz jako wiceprezes ds. badań naukowych w International Drug Development Institute (Belgia).

Głównymi obszarami jego zainteresowań naukowych są metodologia prób klinicznych, meta-analizy prób klinicznych, walidacja zastępczych kryteriów oceny skuteczności leczenia, analiza przeżycia, oraz analiza danych “omicznych” (genomicznych itp.).

Jest współautorem, wraz z Andrzejem Gałeckim, książki

Linear Mixed-effects Models Using R. A Step-by-step Approach.



BioinfoRmatyka nowotworów

Ewa Szczurek

Uniwersytet Warszawski

Kontakt: szczurek@mimuw.edu.pl

Rak to choroba genomu. DNA komórek rakowych charakteryzują liczne alteracje, o przytaczającej złożoności i różnorodności. W referacie przedstawię trzy podejścia analizy tak skomplikowanych danych genetycznych z komórek nowotworowych i wyciągania z nich konkretnych wniosków o mechanizmach tej choroby.

Wszystkie trzy: muex, SurvLRT i lem, zostały zaimplementowane w R.

Bio

Ewa Szczurek jest adiunktem w Instytucie Informatyki na wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego. Posiada dwa tytuły magistra informatyki: Uniwersytetu w Uppsali, w Szwecji, i Uniwersytetu Warszawskiego. Studia doktoranckie w dziedzinie bioinformatyki ukończyła w Instytucie Maxa Plancka w Berlinie. Odbyła dwa staże podoktorskie, jeden w Berlinie, a drugi na ETH w Zurychu. W swej pracy dydaktycznej stara się zainteresować statystyką studentów kierunku Bioinformatyka.

Od sześciu lat zajmuje się tematyką biologii obliczeniowej nowotworów. Obecnie naukowo rozgryza tematykę powstawania przerzutów w raku, a także koleżeńskie stosunki genów aktywnych w tej chorobie. Wszystkie tworzone modele programuje w R.



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.

Bio

Wit Jakuczun to założyciel i współwłaściciel firmy doradczej WLOG Solutions będącej strategicznym partnerem we wdrażaniu rozwiązań analitycznych dużej skali w oparciu o środowisko R. W firmie jest odpowiedzialny za tłumaczenie potrzeb biznesowych klientów na język matematyki.

Prowadził projekty dla wielu branż: bankowość, energetyka, gaz, marketing, logistyka, farmacja, retail, telekomunikacja, ubezpieczenia. W ramach tych projektów tworzył i wdrażał rozwiązania wielkiej skali w środowisku R (i nie tylko) wykorzystujące modele predykcyjne, optymalizacyjne i symulacyjne.



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 wstydu 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.

Bio

Artur Suchwałko ma dwudziestoletnie doświadczenie w projektach analitycznych. Pracował dla różnych firm, od start-upów po międzynarodowe korporacje i w różnych rolach, od pracownika, przez konsultanta, po właściciela. Jest doświadczonym programistą oraz menedżerem projektów. Przez kilkanaście lat pracy statystyka w banku zajmował się głównie budową modeli predykcyjnych i tworzeniem oprogramowania do ich budowy. W tym samym czasie został doktorem matematyki i napisał kilkanaście prac naukowych. Od pięciu lat rozwija z sukcesem swoją firmę QuantUp, zajmującą się analizą danych, modelowaniem statystycznym i tworzeniem oprogramowania oraz szkoleniami z tych dziedzin.



Przeprowadził przynajmniej kilkadziesiąt projektów i kilka tysięcy godzin komercyjnych szkoleń z głębszej analityki biznesowej. Jest współwłaścicielem, Vice CEO i CSO szwedzkiej firmy bioinformatycznej MedicWave. Od kilkunastu lat wykorzystuje w biznesie darmowe oprogramowanie (głównie R) i promuje jego używanie. Jest fanem R i współautorem wydanej w PWN książki o prognozowaniu w R.

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.

Bio

Małgorzata Bogdan uzyskała dyplom magistra z matematyki (1992) i doktora nauk matematycznych (1996, specjalność statystyka matematyczna) na Politechnice Wrocławskiej i habilitację z nauk technicznych (2009, specjalność informatyka) w Instytucie Podstaw Informatyki Polskiej Akademii Nauk.

Pracuje jako profesor nadzwyczajny w Instytucie Matematyki Uniwersytetu Wrocławskiego. Głównym obszarem jej zainteresowań naukowych jest redukcja wymiaru w dużych zbiorach danych i zastosowania do analizy danych genetycznych.

Współautorka wraz z Florianem Frommletem i Davidem Ramseyem, książki *Phenotypes and Genotypes. Search for influential genes*.



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.

Bio

Michał Ramsza uzyskał stopień magistra w zastosowaniach matematyki (1996) na Uniwersytecie Warszawskim, stopień doktora nauk ekonomicznych (2000) oraz stopień doktora habilitowanego (2010) w Szkole Głównej Handlowej w Warszawie.

Pracował w różnych bankach jako analityk, dyrektor Departamentu Analiz Rynkowych w KNF, ale również jako research fellow w University College London. Współpracuje z Instytutem Badań Strukturalnych.

Obecnie profesor ekonomii matematycznej w Szkole Głównej Handlowej w Warszawie oraz uczestnik w projektach komercyjnych związanych z analizą danych dla firm z różnych sektorów. Jego główne zainteresowania związane są z teorią gier oraz analizą systemów złożonych.



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.

Bio

Dr hab. Maciej Eder, profesor Uniwersytetu Pedagogicznego w Krakowie, od kilkunastu lat pracownik Instytutu Języka Polskiego PAN, od roku jego dyrektor. Doktorat (z literatury XVII wieku) obronił w 2005 na Uniwersytecie Wrocławskim, habilitację (z językoznawstwa kwantytatywnego) w 2014 na Uniwersytecie Pedagogicznym.

Zajmuje się analizą danych językowych, w tym modelowaniem zmian w języku staro- i średniopolskim, przede wszystkim zaś stylometrią, czyli kwantytatywną analizą cech językowych, dzięki którym jesteśmy w stanie np. rozpoznać autorstwo anonimowego tekstu. Swoje prace poświęca głównie testowaniu metod wielowymiarowych na materiale różnych języków: polskim, angielskim, łacińskim, starogreckim etc. Jest m.in. autorem rozprawy na temat "Kroniki polskiej" ("Chronica Polonorum") autorstwa tzw. Galla Anonima, w której weryfikuje hipotezę o weneckim pochodzeniu Galla.



W swojej pracy posługuję się programem R. Jest pomysłodawcą i głównym autorem pakietu "stylo", zawierającego kilkadziesiąt funkcji do analizy tekstu za pomocą różnych metod stylometrycznych. Wielokrotnie prowadził warsztaty analizy tekstu przy użyciu R w różnych ośrodkach akademickich, m.in. w Lipsku, Victorii, Getyndze, Amsterdamie, Frankfurcie, Padwie, Budapeszcie, Edynburgu. Działa aktywnie w środowisku humanistyki cyfrowej (Digital Humanities), w którym szerzy, z lepszym lub gorszym skutkiem, idee analizy statystycznej w zastosowaniu do danych humanistycznych.

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 niometrycznym. W referacie przedstawione zostaną takie wykresy jak: wykres mozaikowy, sitkowy, asocjacji, dwuwarstwowy, czteropolowy, czy też cpcp oraz rmp.

Bio

Justyna Brzezińska w 2000 roku rozpoczęła studia na Uniwersytecie Ekonomicznym w Katowicach na specjalności Statystyka i Ekonometria na Wydziale Zarządzania. W 2005 roku obroniła pracę magisterską z zakresu skalowania wielowymiarowego uzyskując tytuł zawodowy magistra. Po studiach odbyła liczne staże zagraniczne m. in. w Belgii, Malezji oraz Brazylii. W latach 2008-2012 była uczestnikiem Studiów Doktoranckich na Uniwersytecie Ekonomicznym w Katowicach. W 2012 roku wzięła udział w szkoleniach organizowanych w ramach programu LLP- Erasmus, które odbyła na Uniwersytecie Technicznym w Wilnie i Europa-Universität Viadrina we Frankfurcie nad Odrą. W 2014 roku obroniła prace doktorską pt.: „Modele logarytmiczno-liniowe i ich zastosowanie analizie zjawisk ekonomicznych” pod kierunkiem prof. dr hab. Eugeniusza Gatkara na Wydziale Zarządzania Uniwersytetu Ekonomicznego w Katowicach uzyskując stopień doktora nauk ekonomicznych w dyscyplinie ekonomia. Rozprawa ta uzyskała nagrodę Rektora Uniwersytetu Ekonomicznego w Katowicach. W latach 2008-2012 pełniła funkcję sekretarza naukowego dziekana Wydziału Zarządzania Uniwersytetu Ekonomicznego w Katowicach. Od 2012 roku jest asystentem w Katedrze Analiz Gospodarczych i Finansowych na Wydziale Finansów i Ubezpieczeń Uniwersytetu Ekonomicznego w Katowicach. Kierowała grantem Narodowego Centrum Nauki pn. Modele logarytmiczno-liniowe w analizie danych jakościowych, prowadziła szkolenia z zakresu statystyki, a także gościnne wykłady naukowe z zakresu statystyki wielowymiarowej w uczelniach zagranicznych (Università degli Studi di Perugia, Università di Bologna, Università degli Studi di Milano). W 2015 była uczestnikiem szkolenia Tranmsformation.doc w University of Alberta w Kanadzie zorganizowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego. Jest autorką ponad trzydziestu publikacji naukowych, głównie w języku angielskim w prestiżowych czasopismach zagranicznych i krajowych, a także jednej monografii naukowej pt.: „Analiza logarytmiczno-liniowa. Teoria i zastosowania z wykorzystaniem programu R”. Jej artykuły były trzykrotnie nagradzane na konferencjach międzynarodowych. Jej zainteresowania naukowe obejmują: statystyczną analizę wielowymiarową, analizę danych jakościowych w szczególności modele logarytmiczno-liniowe, analizę klas ukrytych oraz modele teorii odpowiedzi na pozycje testowe (modele IRT).



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.

Bio

Anna Wróblewska otrzymała doktorat w dziedzinie informatyki w 2008 roku na Wydziale Elektroniki i Technik Informacyjnych Politechniki. Pracuje jako adiunkt na Politechnice Warszawskiej, wcześniej w Instytucie Informatyki Politechniki Warszawskiej, obecnie na Wydziale Matematyki i Nauk Informacyjnych, w tym prowadzi zajęcia na kierunku Informatyka na nowej specjalności Przetwarzanie i Analiza Danych (Data Science). Prowadzi projekty studenckie, prace magisterskie i inżynierskie, wykłady z zakresu eksploracji danych tekstowych i uczenia maszynowego.



Posiada kilkuletnie doświadczenie w projektowaniu inteligentnych systemów analizy i opisu semantycznego danych nabity w środowisku komercyjnym i naukowym. Obecnie od ponad 3 lat pracuje na stanowisku ekspert analizy danych (Data Scientist) w firmie Allegro – największym portalu e-commerce w Europie Wschodniej, gdzie zajmuje się metodami analizy danych także tekstowych i obrazowych.

Ponadto jest autorką ponad 35 publikacji w polskich i międzynarodowych czasopismach i materiałach. Jej zainteresowania naukowe koncentrują się wokół uczenia maszynowego w praktycznych zastosowaniach, w tym przede wszystkim semantycznego rozumienia danych: tekstu i obrazu, budowania ontologii, wyszukiwania semantycznego, systemów rekomendacyjnych.

Prywatnie jest szczęśliwą żoną i mamą dzieci w wieku od przedszkolaka do nastolatka.

Show Me Your Model

Przemysław Biecek

MI²

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*)

Bio

Dr hab. inż. Przemysław Biecek, prof. PW od kilkunastu lat pracuje nad metodami modelowania wysokowymiarowych danych dużego wolumenu. W roku 2003 ukończył studia magisterskie w specjalnościach “Inżynieria Oprogramowania” i “Statystyka Matematyczna” na Politechnice Wrocławskiej, w roku 2007 obronił doktorat, a w roku 2013 obronił habilitację w obszarze Statystyka Medyczna.



Aktywnie współpracuje z biznesem, wybrane większe projekty: z działem badawczym firmy Netezza przy zanurzaniu natywnego rozproszonego przetwarzania danych wewnątrz hurtowni danych; z działem badawczym IBM przy analizie i wizualizacji danych z sieci społecznościach; z działem badawczym iQor Polska przy profilowaniu zachowań konsumenckich; z działem badawczym OECD przy badaniu rozkładu kapitału naukowego międzynarodowej młodzieży.

W pracy naukowej interesuje się technikami modelowania wysokowymiarowych danych (testowania zbioru hipotez oraz wyboru modelu), jak i wizualizacją danych oraz aplikacjami w medycynie i genomice. Jest autorem lub współautorem 55 publikacji z listy JCR. Wiele z ostatnich publikacji jest poświęconych analizie danych onkologicznych.

Jest autorem trzech popularnych podręczników akademickich poświęconych programowi R (Przewodnik po programie R, GiS 2017), analizie danych (Modele liniowe i mieszane, PWN 2015) i wizualizacji danych (Odkrywać! Ujawniać! Objaśniać! 2016). W ramach działalności popularyzującej naukę współorganizuje liczne konferencje i spotkania poświęcone programowi R (Why R? 2017, UseR 2017, eRum 2016, SER 2014-2017, WZUR 2008-2012).

Pracuje również nad projektem pozaakademickiej edukacji statystycznej (projekt Beta i Bit).

Sesje



pwc

PwC to czwarta najsilniejsza marka świata według badania Brand Finance, a jednocześnie zwycięzca w 3 rankingach Warsaw Business Journal: doradztwa biznesowego, podatkowego oraz usług audytowo-księgowych.

Firma PwC kieruje się w swojej działalności trzema głównymi wartościami: jakością i doskonałością, pracą zespołową oraz przywództwem. Świadczy usługi z zakresu auditu, doradztwa biznesowego, podatkowego i prawnego, jak również w obszarze digital transformation. Jest obecna w 157 krajach zatrudniając ponad 233 tysięcy pracowników na całym świecie.

W Polsce PwC zatrudnia zespół blisko 3 500 pracowników w ośmiu miastach: w Gdańsku, Katowicach, Krakowie, Łodzi, Poznaniu, Rzeszowie, Wrocławiu i Warszawie.

Logistyka i Supply Chain Management z wykorzystaniem R

Paweł Pitera

PwC

Na prezentacji przedstawione zostanie zagadnienie biznesowe związane z optymalizacją sieci dostaw i optymalnym wyborem punktów koncentracyjnych (depotów), jego symulacja i rozwiązywanie w R jak i w połączeniu z innymi narzędziami powszechnie wykorzystywany do problemów typu Supply Chain Management.

Rozpraszanie wielowątkowe na przykładzie analizy transakcji w Retailu

Michał Cisek, Rafał Kobiela

PwC

Podczas wystąpienia przedstawiony zostanie typowy problem analizy transakcji wpływu promocji na volumen i marżę sprzedaży w długim okresie czasu. Poruszone zostaną zagadnienia up-sellu, cross-sellu, cherry-pickingu i kanibalizacji sprzedaży. Pokażemy także jak można użyć R'a by podejść do problemu analizy paragonów, który wkracza w obszar Big Data, i jak w prosty sposób można zrównoleglić obliczenia na wiele maszyn połączonych w klaster obliczeniowy na przykładzie infrastruktury konwergentnej.

Wykorzystanie R w rozwiązywaniu problemów biznesowych w branży finansowej

Katarzyna Lenczewska, Ewelina Chmura

PwC

Na prezentacji przedstawione zostaną przykłady użycia analizy danych i budowy modeli predykcyjnych do rozwiązywania najbardziej popularnych problemów i wyzwań z branży finansowej. Pokażemy jak tworzyć DataMarty na potrzeby analityczne w postaci relacyjnych baz danych i jak wpleść modele R'owe na nich bazujące w środowisku produkcyjne.

Tidyverse

Jak zbudowaliśmy aplikację Shiny dedykowaną 700 użytkownikom?

Olga Mierzwa-Sulima

Apppsilon

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 skompilowanymi zależnościami zmiennych reaktywnych. Zademonstrujemy aplikację i powiemy jak jej wdrożenie przełożyło się na biznes klienta.

Jak wygrać więcej w lotto z R?

Błażej Kochański

Politechnika Gdańsk, 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.

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 blogdownem

Natalia Potocka

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 zaleta 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 szuchtaawdanych.pl

Modelowanie Ryzyka (UBS)



UBS jest jedną z największych instytucji finansowych na świecie, z ponad 150-letnią historią. Działalność firmy koncentruje się na trzech głównych obszarach: zarządzaniu majątkiem, zarządzaniu aktywami oraz bankowości inwestycyjnej. Zatrudniamy ponad 60000 osób w ponad 50-ciu krajach. Główna siedziba UBS znajduje się w Szwajcarii, natomiast w biurach w Krakowie i we Wrocławiu pracownicy współpracują w ramach zespołów zlokalizowanych w różnych regionach Europy i świata.

UBS tworzy środowisko ludzi zafascynowanych modelowaniem, którzy chętnie dzielą się swoją wiedzą, a język R jest szeroko używany wewnątrz UBS w obszarze modelowania ryzyka.

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 przedstawienie 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 między 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 przywiążują 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 modelowania 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

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 wewnętrz 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 wewnętrz 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. Zaproponujemy 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.

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 – szeregow 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.

Planowanie pojemności i wydajności w celu przeciwdziałania awariom

Robert Bigos

Pracował dla firm 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 wszysko kiedyś ulegnie awarii. Coraz częściej spotykamy się ze zjawiskiem “capacity rolling disaster” czyli awariami w których pierwotna przyczyna tylko inicjujełańczych zdarzeń. Duże systemy instrumentacyjne (logi/metryki) zbierają rocznie TBy danych, w rozdrobniono 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.

FSelectorRcpp – wolna od Java/Weka implementacja pakietu FSelector

Krzysztof Słomczyński

BD Polska

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 sentymenu przy użyciu bibliotek Microsoft

Łukasz Grala

TIDK - Data Scientist as a Service

Analiza sentymenu jest powszechnym wyzwaniem wielu dużych organizacji. Analizujemy treści poczty elektronicznej, dyskusji na forach, tekstów z komunikatorów, czy też różnorodnych portali społecznościowych. Analiza 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 dostarczany przez firmę Microsoft. Rozwiążanie dostępne zarówno w środowisku Windows, jak i Linux, z poziomu SQL Server, hurtowni danych Teradata, czy też rozwiązań pracujących na HADOOP czy Spark

Biostatystyka i Edukacja

Identyfikacja i analiza barier przeciwko introgresji pomiędzy 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.

MLEpResso – NGS, Metylacja, Expresja, R i sporo kawy

Alicja Gosiewska, Aleksandra Dąbrowska

Politechnika Warszawska MI²

W naszym wystąpieniu przedstawimy MLExpResso - pakiet służący do jednoczesnej analizy i wizualizacji danych dotyczących ekspresji genów oraz metylacji DNA. Kluczowymi funkcjami pakietu są:

- Identyfikacja regionów o zróżnicowanej metylacji (DMR),
- Identyfikacja genów o zmienionej ekspresji,
- Identyfikacja regionów ze zmianami w ekspresji i metylacji,
- Wizualizacja zidentyfikowanych regionów.

Wspólne modelowanie i wizualizacja ekspresji oraz metylacji zwiększa interpretowalność zidentyfikowanych sygnałów.

survminer - wykresy analizy przeżycia pełne informacji i elegancji

Marcin Kosiński

Gradient Metrics

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.

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 zgłoszenia 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.

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 oczekiwaniemi, 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, jak i na podejście zespołu do nowych problemów.

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

Anna Skrzypdło

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 procesów 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

Dla kogo ?

Dla wszystkich zaangażowanych w tworzenie społeczności R w Polsce. Zarówno dla osób, które takie grupy prowadzą, jak również dla tych którzy chcą utworzyć społeczność entuzjastów R w swojej okolicy :)

Dlaczego ?

Grupy pasjonatów odgrywają ogólną rolę w popularyzacji R. Celem każdego ze spotkań jest wymiana wiedzy i doświadczeń pomiędzy obecnymi i nowymi użytkownikami, kształcenie oraz networking. W Polsce aktywnie działa 6 grup. W trakcie sesji chcielibyśmy pomóc w znalezieniu inspiracji do dalszego działania oraz ułatwić wymianę doświadczeń związanych z organizacją społeczności skupiających entuzjastów R w Polsce.

Czego można się spodziewać ?

Żywiołowej dyskusji pomiędzy przedstawicielami istniejących społeczności R w Polsce, pod kontrolą charyzmatycznego prowadzącego :)

Panel dyskusyjny: Edukacja statystyki

Jak uczyć analizy danych?

Z tym pytaniem zmierzą się świetni dydaktycy z najlepszych polskich uczelni.

Panel składać się będzie z dwóch części. W pierwszej każdy z zaproszonych panelistów będzie miał okazję odpowiedzieć na trzy pytania. Druga część będzie oparta o dyskusje z publicznością i wolne wypowiedzi.

Paneliści:

- Małgorzata Bogdan (Uniwersytet Wrocławski),
- Tomasz Burzykowski (Hasselt University/Politechnika Warszawska),
- Przemysław Grzegorzewski (Politechnika Warszawska),
- Paweł Teisseyre (Politechnika Warszawska),
- Michał Ramsza (Szkoła Główna Handlowa).

Panel poprowadzi Przemysław Biecek (Grupa MI2).

Lightning Talks

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.

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 (nietendencjonalna 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: labds, 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.

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, 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 problemem, wzięliśmy pod lupę dane z automatycznych liczników, które od kilku lat zliczają ruch rowerowy w kilkunastu miejscowościach 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.

Visualizing statistical analysis methods on the data from The Cancer Genome Atlas Study available in the RTCGA Family of R packages

Marcin Kosiński, Witold Chodor, Przemysław Biecek

The following poster presents RTCGA: a family of R packages with data from The Cancer Genome Atlas Project (TCGA) study. TCGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. We converted selected datasets from this study into few separate packages that are hosted on Bioconductor. These R packages make selected datasets easier to access and manage. Data sets in RTCGA.data packages are large and cover complex relations between clinical outcomes and genetic background.

These packages will be useful for at least three audiences: biostatisticians that work with cancer data; researchers that are working on large scale algorithms, for them RTGCA data will be a perfect blasting site; teachers that are presenting data analysis method on real data problems. In this poster we present applications of the statistical modeling methods to the cancer's subtypes basing on the available cancer data of over 8,000 patients.

Warsztaty

Analiza danych sondażowych w R



Dariusz Szklarczyk, Agnieszka Otręba-Szklarczyk

Fundacja Rozwoju Badań Społecznych

Opis warsztatu

Wśród badaczy społecznych korzystających używającymi 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 najpowszechniejsze 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 sporządzania 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, polityologów, 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, warstwowej) i ważenie próby (klasyczne oraz rake weights).

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?
2. Pakiet „survey” - jego możliwości i ograniczenia.
3. Definiowanie typowych złożonych schematów doboru próby w pakiecie „survey”.
4. Estymacja typowych statystyk opisowych.
5. Przerwa.
6. Wizualizacja danych przy pomocy pakietów „survey” i „ggplot2”.
7. Regresja liniowa i uogólniona regresja liniowa.
8. Poststratyfikacja i techniki pokrewne (jeśli ktoś jest zainteresowany, może rzucić okiem na tą prezentację).

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

Pakiet rvest, czyli web scrapingu wybrane przypadki



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).

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 przeglądarki. W pierwszej kolejności zajmiemy się krótko scrapowaniem stron statycznych, niekorzystających z javascriptu jako prezentacja, że tradycyjne scrapowanie jest również możliwe, powiemy jednak czemu jest to nieefektywne. Następnie przejdziemy do części głównej, tj. scrapowania stron korzystających 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 PhantomJS.

MicrosoftML

- State of the art Machine Learning Microsoft



Lukasz 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

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. Możliwość 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

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

Kombajn do uczenia maszynowego

- MLR w praktyce



Paweł Zawistowski

Politechnika Warszawska, Wydział EiT, AdForm

Opis warsztatu

Kiedy coraz powszechniejsze staje się stosowanie mniej lub bardziej skomplikowanych modeli statystycznych, a liczba pakietów R'a z nowymi metodami i algorytmami ciągle wzrasta - dobrze mieć w zanadrzu narzędzie, które spina wszystkie etapy tworzenia modelu w jedną całość. Pakiet MLR jest takim właśnie "kombajnem", który może w znacznym stopniu ułatwić nam pracę.

W ramach warsztatu zobaczymy jakie możliwości daje MLR przy tworzeniu różnego rodzaju modeli - przejdziemy przy tym kompletną ścieżkę, od wstępnego przygotowania danych, przez wybór odpowiedniej metody, strojenie parametrów, aż po diagnostykę i wizualizacje wyników.

Plan warsztatu

1. Omówienie możliwości pakietu MLR, przygotowanie środowiska.
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.

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)

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

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

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

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 korespondencji (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

mirt: skalowanie odpowiedzi lepsze niż PCA



Piotr Migdał

deeepsense.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 równe 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

Social Network Analysis w R



Michał Wojtasiewicz

Instytut Podstaw Informatyki PAN

Opis warsztatu

Celem warsztatu jest zapoznanie 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 najprężej 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 restrykcją braku przestoju.

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)

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 komplikowanym 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++.

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 kolumn 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.

Sponsorzy brązowi



Sages jest firmą szkoleniową i doradczą, świadczącą usługi w zakresie IT zarówno dla firm z branży informatycznej, jak i firm z innych obszarów działalności, dla których wytwarzanie systemów jest koniecznym warunkiem rozwoju.

Naszą misją jest wspieranie procesu projektowania, wytwarzania i wdrażania oprogramowania w przedsiębiorstwach i instytucjach publicznych, aby zminimalizować koszty, czas i ryzyko, związane z tymi przedsięwzięciami. Naszą misję realizujemy poprzez podnoszenie kompetencji kadr w postaci szkoleń, coachingu i outsourcingu kompetencji, a także poprzez świadczenie wysoce specjalizowanych usług doradczych.

Wizją Sages jest łączenie wiedzy pochodzącej ze środowisk akademickich i doświadczenia ekspertów z obszaru praktycznych zastosowań w biznesie, aby budować kompetencje kadr w Polsce, a w szczególności przyczynić się do wzrostu efektywności i innowacyjności krajowej branży IT.



Pearson w Poznaniu to ponad 160-osobowy, pełen pasji i zaangażowania zespół zajmujący się tworzeniem platform i materiałów e-learningowych wspierających naukę języka angielskiego. Jesteśmy częścią grupy Pearson – międzynarodowej organizacji z ponad 200-letnią historią, co daje nam możliwość pracy nad produktami o globalnym zasięgu, jak i realnego kształtowania przyszłości edukacji na świecie!

Aktualnie mieścimy się w Poznaniu w Nobel Tower, w centrum miasta. Nasz zespół składa się z działów: IT, Design Team, Content Creation, Efficacy and Research, Digital Product Support oraz jednostek wpierających (Administracja, HR, Finanse).

Indeks nazwisk

- Ćwiakowski Piotr, 67
Żółtak Tomasz, 57
- Alekseichenko Vladimir, 64
- Biecek Przemysław, 29, 53
Bielski Włodzimierz, 49
Bigos Robert, 40
Bogdan Małgorzata, 24
Bogucki Mikołaj, 44
Brzezińska Justyna, 27
Burzykowski Tomasz, 20
- Chmura Damian, 48
Chmura Ewelina, 33
Chodor Witold, 53
Cisek Michał, 33
Czernecki Bartosz, 49
Czortek Patryk, 52
- Dąbrowska Aleksandra, 42
Detko Dawid, 61
Dyderski Marcin K., 39, 68
- Eder Maciej, 26
- Gosiewska Alicja, 42
Grala Łukasz, 40, 60
- Jędrzejewski Krzysztof, 45
Jagodziński Andrzej M., 39
Jakuczun Wit, 22
- Kobiela Rafał, 33
Kochański Błażej, 34
Kosiński Marcin, 43, 53
Kowalczyk Dorota, 37
- Lenczewska Katarzyna, 33
Lodzikowski Kacper, 49
- Młodożeniec Marek, 35
Maj Michał, 62
Martsenyuk Vasyl, 45
- Mazurek Magdalena, 59
Melcer Tomasz, 43
Mierzwa-Sulima Olga, 34
Migdał Piotr, 69
- Ocalewicz Piotr, 66
Ochotny Stanisław, 36
Oleś Andrzej, 43
Olszewski Mikołaj, 44
Otmanowski Mateusz, 48
Otręba-Szkłarczyk Agnieszka, 56
- Pankowska Emilia, 45
Pawlowska Monika, 53
Pawlik Łukasz, 52
Pitera Paweł, 33
Potocka Natalia, 35
- Ramsza Michał, 25
Ryciak Norbert, 71
- Sękiewicz Bartosz, 58
Słomczyński Krzysztof, 40
Samonil Pavel, 52
Skrzydło Anna, 44
Sobczyk Piotr, 35
Staniak Mateusz, 50
Stankiewicz Krzysztof, 42
Suchwałko Artur, 23
Szczurek Ewa, 21
Szklarczyk Dariusz, 56
- Tartanus Bartłomiej, 73
Teisseyre Paweł, 65
- Wójcik Piotr, 38
Wojtasiewicz Michał, 70
Wróbel Adam, 37
Wróblewska Anna, 28
- Zółtak Tomasz, 45
Zawadzki Zygmunt, 72
Zawistowski Paweł, 63

Sponsor Platynowy



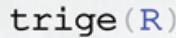
Sponsor Złoty



Sponsorzy Brązowi



Partnerzy Społecznościowi



Współorganizator

