Conference Report: Why R? 2019

by Michał Burdukiewicz, Filip Pietluch, Jarosław Chilimoniuk, Katarzyna Sidorczuk, Dominik Rafacz, Leon Eyrich Jessen, Stefan Rödiger, Marcin Kosiński and Piotr Wójcik

Figure 1: *Why R?* 2019 conference banner used for social media promotion.

Why R? 2019 conference

Why R? conferences have been the hallmark of the Why R? Foundation (whyr.pl). Our goal has been to establish a series of international R-related events in Poland. After three years, we are happy to announce that our main event, the Why R? conference, has become one of the largest annual R conferences in Central Europe.

Why R? 2019 was the third part of Why R? conference event. After the last edition that was held in Wrocław (**?**), our conference has returned to Warsaw. A total of approximately 300 people from 20 countries attended the main conference event. The event took place from 26th to 29th September 2019 and was co-organised by **the Faculty of Economic Sciences of the University of Warsaw** (wne.uw.edu.pl/en/), a leading academic institution in Poland, having important achievements in quantitative methods and data science. We received major support from **ML in PL Society** (mlinpl.org), a group of young researchers, aiming to promote machine learning events in Poland, who shared their resources and experience to make the conference more accessible.

For the first time, this year the conference featured a language-agnostic data visualizations hackathon (whyr.pl/2019/hackathon). Such an event gives the Why R? community a chance to exchange experience and inspirations with the users of any other languages and tools.

Participants

In spite of the fact that *Why R?* events are aimed at experienced data science practitioners, each conference gathers a high percentage of students (around 30%). Our participants have very diverse scientific backgrounds, where mathematics (mainly statistics) and computer science are the most common. All of them have jobs related to data science, including professional R developers (programmers), data engineers, machine learning practitioners and business analysts. One of the key advantages of *Why R?* is that it gathers participants

both from academia and the industry.

Conference program

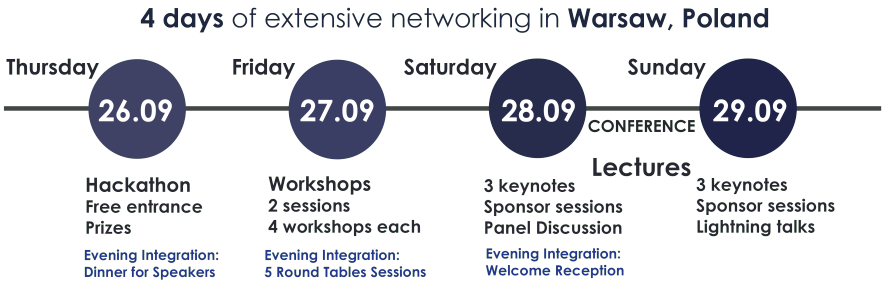

Figure 2: *Why R?* 2019 conference programme.

The format of the conference was aimed at exposing participants to recent developments in the R language as well as a wide range of application examples. The event consisted of workshops, invited keynote talks, field-specific series of talks, lightning-talks, special interest groups and a full-day data visualizations hackathon. It offered extensive networking opportunities. The welcome party was held at the conference venue on the first day of lectures. In addition, many informal gatherings were organised during each conference day, as the event took place close to the Old Town.

To sum up, *Why R?* 2019 consisted of: one day of hackathon (60 attendees), one day of workshops (150 attendees), one evening of round tables, two days of lectures (250 attendees) and one evening Welcome paRty (100 attendees). In 2019 we hosted a total of 315 unique attendees. During lectures there were carried out: 6 keynote talks, 42 regular talks and 14 lightning talks. Below you can find the conference agenda.

	Saturday 28.09		Sunday 29.09	
08:00-08:30	Registration		Registration	
08:30-09:00	Welcome Session			
09:00:09:45	**Marvin Wright** A		**Steph Locke** A	
09:45-10:00	Coffee Break		Coffee Break	
10:00-11:20	**Shiny** A	**Modelling 1** B	**Philosophy** A	**Modelling 2** B
11:20-11:35	Coffee Break		Coffee Break	
11:35-12:35	**Scoring** A	**API** B	**XAI** A	**EDA** B
12:35-14:05	Lunch		Lunch	
14:05-14:50	**Jakub Nowosad** A		**Wit Jakuczun** A	
14:50-15:05	Coffee Break		Coffee Break	
15:05-16:05	**GEO** A	**BIO 1** B	**Lightning 1** A	**Vision 1** B
16:05:16:20	Coffee Break		Coffee Break	
16:20-17:20	**Business** A	**BIO 2** B	**Lightning 2** A	**Vision 2** B
17:20-17:35	Coffee Break		Coffee Break	
17:35-18:20	**Sigrid Keydana** A		**Paula Brito** A	
18:20-18:40	Coffee Break		Closing Remarks	

Figure 3: *Why R?* 2019 conference agenda.

Materials from the conference are available on GitHub and YouTube:
- abstracts github.com/WhyR2019/abstracts,
- presentations github.com/WhyR2019/presentations
- videos whyr.pl/youtube/

Data Visualizations Hackathon

On the day before the conference we organized the free Data Visualizations Hackathon. It was a great opportunity for networking and exchange of experiences between data scientists that use different programming languages. The challenge was based on the data from Google Places API, which allows to search for places in a particular area. Thanks to this API we gathered data related to places in Warsaw, their working hours and occupancy. Based on this source of data participants, divided into 10 teams, were asked to prepare useful business application powered data visualizations solutions and techniques.

Pre-meetings

In 2019, *Why R?* 2019 was preceded by fourteen pre-meetings in eight countries. The purpose of those meetings was to provide the space for professional networking and knowledge exchange for practitioners and students, from the area of statistical machine learning, programming, optimization and data science. The *Why R?* Foundation supported organisation of pre-meetings financially and/or by sending speakers.

The organisation of pre-meetings would not be possible without the wonderful support of local R communities. Aside from the promotion of *Why R?* we had a great opportunity to interact with other R enthusiasts.

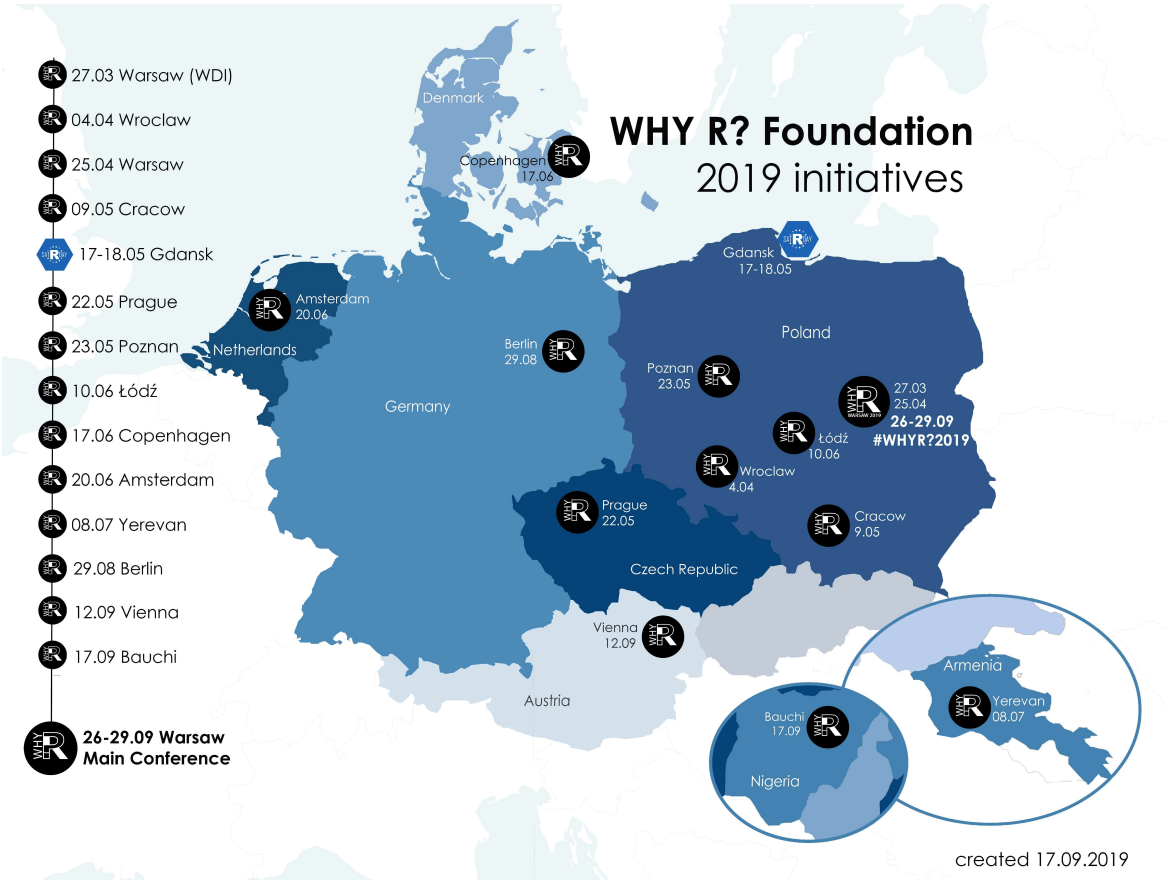

Figure 4: Locations and dates of the main *Why R?* 2019 conference and *Why R?*-branded pre-meetings.

Workshops

	Friday 27.09				
	Room B	Room C	Room E	Room A103	Room A203
08:00-09:00	Registration				
09:00:10:30	mlr3 and drake	Generalized Additive Models	data.table	Shiny basics	auditor + DALEX
10:30-11:00	Coffee Break				
11:00-12:30					
12:30-14:00	Lunch				
14:00-15:30	Deep Learning in R (with Keras)		Speeding up R wih C++ (Rcpp)	Basics of spatial data analysis	eXplainable Artificial Intelligence in Business
15:30-16:00	Coffee Break				
16:00-17:30					
17:30-18:00	Coffee Break				

Figure 5: *Why R?* 2019 workshops.

Why R? 2019 conference had a wide portfolio of workshops that are listed below. One can find materials from workshops at this GitHub repository `github.com/WhyR2019/workshops`

- **Introduction to modern Generalized Additive Models in R (with mgcv)** by Matteo Fasiolo (University of Bristol). The assumption of the full-day workshop was firstly to give its participants some theoretical background about GAMs and some practical experience in R and finally to make attendees ready to start applying these models themselves. GAM models are a non-parametric extension of traditional regression model and were proved to be highly useful for both predictive and inferential purposes. Their popularity is based on a good balance between flexibility and interpretability as

well as on the possible application on large datasets. Matteo started with explanation of standard GAMs and related R packages. He explained what an additive model is, how the smooth effects and random effects are introduced. GAM models fitting was accompanied by the explanation of additional diagnostic and model selection tools, and Big Data GAM methods. In the end more recent developments were also described, i.e. quantile GAM models. The practical sessions were based on the **mgcv** (**?**), **qgam** (**?**) and **mgcViz** (**?**) packages.

- **data.table introduction & time-series** by Jan Gorecki (H2O.AI). The workshop was divided into two parts – the first part was devoted to the introduction of `data.table` query concept while the second focused on a particular use case of working with time-series data. In the first part Jan showed syntax similarities and differences between `data.table` and `data.frame` approaches. He used Arun Srinivasan workshops materials from useR!2017, with a few extras: chaining of `data.table` queries, reference semantics, subset of data: `.SD` and R function argument matching. In the second part Jan showed the application of efficient data processing on financial time series data of high-frequency (tick data quotations), including efficient aggregation to OHLC data, calculation of moving averages and using rolling join.

- **Straightforward introduction to Deep Learning in R (with Keras)** by Mikołaj Bogucki and Mikołaj Olszewski (iDash). The workshop started with the explanation of what Deep Learning and Neural Networks are (complex functions) and what components they include (input, output, hidden layers and weights). Then **Keras** (**?**) was presented as a high level library allowing to build neural networks with an easy to use set of commands. The practical example using airBnB data and **Keras** R codes showed all stages of building a Neural Network: (1) defining the structure of the network, (2) defining the way of training (the loss function and the optimizer algorithm), (3) training (together with its visualization), (4) evaluation and (5) prediction. In addition, training, validation, test set division, simple imputation of missing data, using non-linear activation functions and basic feature engineering was shortly explained.

- **auditor + DALEX: a powerful duet for validation and explanation of machine learning models** by Alicja Gosiewska and Tomasz Mikołajczyk (MI2 Data Lab). The aim of the workshop was to familiarize participants with modern methods of model verification and exploration. In the first part Alicja and Tomasz introduced the idea of **DALEX** (**?**) explainers, showing how to use them to assess the performance of a model and explain the model's predictions (including global and local explanations). In the second part they focused on additional functionalities of the **auditor** (**?**) package, showing how the analysis of residuals may be applied to select the best model or even improve models.

- **Black is the new White - using eXplainable Artificial Intelligence in Business** by Marcin Chlebus (Faculty of Economic Sciences, University of Warsaw, Data Juice Lab, Data Donuts). Marcin presented XAI as a possible solution for understanding "Black-box" model complexity and fuzziness. He showed how XAI helps in stability and sensitivity analysis, prediction quality assessment and identification of decision drivers. The use cases showing the application of XAI in cross-sell marketing campaigns and risk management were presented. With the use of step by step analysis, it was shown that XAI is a set of tools enabling application of "black box" models in many business industries through in-depth understanding of advanced machine learning modelling.

- **Shiny Basics** by Theo Roe (Jumping Rivers). This workshop was intended as a quick introduction to creating interactive visualisations of data using **shiny**. Theo started with some basic examples of using **rmarkdown** and **htmlwidgets**, then showed input and output bindings to interact with R data structures and using inputs to render output tables and graphs. In the end, Theo showed how to create own page layouts using **shiny** and **shinydashboard** and input and output "slots".

- **Speeding up R wih C++ (Rcpp) – from basics to more advanced applications** by Piotr Wójcik (Faculty of Economic Sciences, University of Warsaw, Data Science Lab). Piotr discussed various aspects of **Rcpp** that helps to easily replace the R code with often significantly faster counterparts in C++. Writing R functions in C++ was explained, starting from simple examples with the focus on similarities and differences between R and C++ syntax. Then Piotr at first explained writing loops and recursive calls in C++, using Rcpp sugar and secondly presented how to store C++ code in `*.cpp` files, using Standard Template Library, iterators, algorithms and range-based loops. In the end complex input/output objects (S3 and S4) were discussed.

- **Machine Learning Pipelines and Reproducible Research with mlr3 and drake** by Jakob Richter (TU Dortmund University) and Patrick Schratz (LMU Munich). The workshop was divided into two parts – the former introduced the new **mlr3** package (**?**) framework (the successor of the **mlr** package) while the latter presented a brief overview of the **drake** package (**?**) in R. In the first part Jakob and Patrick explained the philosophy and ingredients of **mlr3** package. They presented how to define the data and the target variable, using learners provided by **mlr3**, set and tune hyperparameters, make predictions and evaluate their performance, including resampling techniques and comparing multiple learners. The practical example showed hyperparameter tuning and training of a random forest classifier on the `iris` dataset. The practical part also involved benchmark analysis of multiple learners, using different hyperparameter ranges on the `iris` and `spam` datasets. A particular emphasis was put on machine learning workflows that might be easily controlled with **mlr3pipelines** package (**?**). In the second part the **drake** package was presented. It helps to set up a reproducible workflow of the project and it easily integrates with the **mlr3** package and its extensions.

- **Basics of spatial data analysis** by Jakub Nowosad (Adam Mickiewicz University, Poznan). The emphasis in this workshop was put on getting started with spatial data analysis. Jakub demonstrated key packages for spatial analysis and making maps, explained spatial data representation in R, using **sf** (**?**), for spatial vector data, and **raster** (**?**) packages. Then he gave a lot of examples of spatial data visualization, using a powerful **tmap** package (**?**), including some vector-raster interactions. In the end, he also showed data manipulation examples with the tidyverse (**dplyr**) approach, in which sf spatial objects are simply special data frames.

Invited talks

The invited talks topics included domain knowledge from statistics, computer science, natural sciences and economics. The speakers list presents as follows:

Marvin Wright

Random forests used to be everywhere, from Microsoft Kinect to meteorology, but their popularity considerably dropped with the advent of deep learning. During his keynote talk at *Why R?* 2019 Marvin R. Wright has shown that random forests still can be used in machine learning routines, making the whole process time- and cost-efficient.

Implementing a real-life machine learning solution is not only about the best performance. Marvin has shown that considering trade-off between performance and costs of the analysis, random forests are still unbeatable. Aside from the methodological background, Marvin has given an overview of random forest implementations in R (**?**).

Marvin is a Postdoc at the Leibniz Institute for Prevention Research and Epidemiology in Bremen, Germany. He is the author of several R packages, including the fastest implementation of random forest in R, ranger. He holds a Ph.D. in Biostatistics from the University of Lübeck, supervised by Andreas Ziegler. In the past, Marvin worked at the University

of Lübeck. He was a visiting researcher at the University of Copenhagen. Also, he spent some time in the automotive and health insurance industries. His main research interests are interpretable machine learning, genetic epidemiology and survival analysis.

Jakub Nowosad

Jakub Nowosad's keynote lecture was a great opportunity to learn about geostatistics. Jakub, a co-author of the Geocomputation with R (**?**), has focused on tools used to solve real-life problems in spatial data analysis.

The growing importance of spatial data stimulates a rapid evolution of geostatistical methods. Jakub, as the active member of #rspatial community, not only presented cutting-edge tools but also gave his unique insight into the future of the spatial data analysis.

Jakub is an assistant professor in the Department of Geoinformation at the Adam Mickiewicz University in Poznan, Poland. His main research is focused on developing and applying spatial methods in order to expand our understanding of processes and patterns in the environment. He has extensive teaching experience in the fields of spatial analysis, geostatistics, statistics, and machine learning.

Sigrid Keydana

We know how accurate are our predictions but do we really know how certain they are? This question has been answered by Sigrid Keydana (RStudio) during her keynote lecture.

Sigrid has presented tfprobability, an interface to TensorFlow Probability, a tool for obtaining uncertainty estimates from deep neural networks. This exciting tool can be extended beyond a classic deep learning framework into complex hierarchical models.

Sigrid is an Applied Researcher at RStudio. She has experience as a psychologist, software developer and data scientist. She is passionate about exploring the borders of deep learning, especially by helping users to apply the power of deep learning in R.

Steph Locke

Machine learning models find their place in almost every area of our life, influencing things as small as the video recommendations on YouTube or as big as the length and severity of a sentence in a criminal procedure. With the growing importance of machine learning, it becomes more and more important to train models while keeping in mind their ethical consequences.

During her keynote talk at *Why R?* 2019, Steph Locke showed us ethical concerns about data science. Apart from pointing out existing issues, she has also presented solutions leading to more fair and transparent machine learning models.

Steph is the founder of a consultancy in the UK. Her talks, blog posts, conferences, and business all have one thing in common – they help people get started with data science. Steph holds the Microsoft MVP award for her community contributions. In her spare time, Steph plays board games with her husband and takes copious pictures of her doggos.

Wit Jakuczun

Wit Jakuczun from WLOG Solutions presented his talk about deploying - How to make R great for machine learning in (not only) Enterprise.

For many years software engineers have put enormous effort to develop best practices to deliver stable and maintainable software. How R users can benefit from this experience? Wit answered this question by going through several concepts and tools that are natural for software engineers but are often undervalued by R users.

Paula Brito

During her keynote lecture at Why R? 2019 Paula Brito has given a unique insight into the world of symbolic data, where data points are represented not as single values, but more complex structures, like sets or intervals (**?**).

A classical paradigm of data science assumes that categorical variables, like gender or educational stage, are represented as the single value per observation. Paula has shown how to utilize her package, MAINT.Data, to model interval data, using its symbolic representation which leads to more accurate and robust models.

Paula is Associate Professor at the Faculty of Economics of the University of Porto, and member of the Artificial Intelligence and Decision Support Research Group (LIAAD) of INESC TEC, Portugal. Her current research focuses on the analysis of multidimensional complex data, known as symbolic data, for which she develops statistical approaches and multivariate analysis methodologies.

Round tables

Round tables are networking-oriented social mixers devoted to connecting people with similar interests. The exact points discussed during the round table and its style depend on the moderators who are shaping out the details, based on the general agenda provided by the *Why R?* organizers. The organizing committee both selects the topics of round tables and invites appropriate moderators.

Diversity in Data Science

This board aims to inspire members of affinity groups to pursue careers in data science. We hope that this platform for networking will reduce the diversity of R community. Moderator: Barbara Sobkowiak (Women in Machine Learning & Data Science Poland).

Career-planning in Data Science

Participants of WhyR will have a chance to learn from more experienced R enthusiasts about their career paths. Moderator: Kamil Kosiński (PwC).

Teaching Data Science

Practitioners will share their experiences in introducing their students to basic and advanced concepts of data science. Moderator: Patrick Schratz (Ludwig Maximilian University of Munich).

Data Visualizations

Discuss data visualizations good practices and approaches to various presentation challenges. Moderator: Michał Burdukiewicz (Warsaw University of Technology).

Ethics in Data Science

With the increased importance of machine learning, we are becoming more and more concerned about the ethics of data science. Moderator: Steph Locke (Locke Data).

Conference organizers

The organizing committee consisted of Klaudia Korniluk, Marcin Kosiński, Michał Burdukiewicz, Jarosław Chilimoniuk, Katarzyna Sidorczuk, Filip Pietluch, Weronika Puchała and Dominik Rafacz.

The quality of the scientific program of the conference was the achievement of Stefan Rödiger (Brandenburg University of Technology Cottbus-Senftenberg), Piotr Wójcik (University of Warsaw) and Bernd Bischl (Ludwig Maximilian University of Munich).

Acknowledgements

We would like to express our gratitude to all our sponsors, the Faculty of Economic Sciences (University of Warsaw), ML in PL Society, local organizers of the pre-meetings and student helpers.

Additional information

***Why R?* 2019 website** http://whyr.pl/2019
Corporate sponsors: PwC Poland, iDash, R Consortium, umping Rivers Ltd., Appsilon Data Science, RStudio, Inc., Analyx®GmbH, Pearson IOKI and WLOG Solutions.

Michał Burdukiewicz
Warsaw University of Technology, Why R? Foundation
Pl. Politechniki 1, 00-661 Warsaw
Poland
michal@whyr.pl

Filip Pietluch
University of Wrocław
Pl. Uniwersytecki 1, 50-137 Wrocław
Poland
fpietluch@gmail.com

Jarosław Chilimoniuk
University of Wrocław
Pl. Uniwersytecki 1, 50-137 Wrocław
Poland
jaroslaw.chilimoniuk@gmail.com

Katarzyna Sidorczuk
University of Wrocław
Pl. Uniwersytecki 1, 50-137 Wrocław
Poland
sidorczuk.katarzyna17@gmail.com

Dominik Rafacz
Warsaw University of Technology
Pl. Politechniki 1, 00-661 Warsaw
Poland
dominikrafacz@gmail.com

Leon Eyrich Jessen
Technical University of Denmark

Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark
Denmark
ljess@dtu.dk

Stefan Rödiger
Brandenburg University of Technology Cottbus–Senftenberg
Universitätsplatz 1, Senftenberg
Germany
ORCiD: 0000-0002-1441-6512
stefan.roediger@b-tu.de

Marcin Kosiński
Gradient Metrics LLC, Why R? Foundation
Warsaw
Poland
marcin@whyr.pl

Piotr Wójcik
University of Warsaw, Data Science Lab
ul. Dluga 44/50, 00-241 Warsaw
Poland
pwojcik@wne.uw.edu.pl