

Краткий экскурс в обработку данных в среде программирования R



ИЭРиЖ

ИНСТИТУТ ЭКОЛОГИИ
РАСТЕНИЙ И ЖИВОТНЫХ

Артём Созонтов

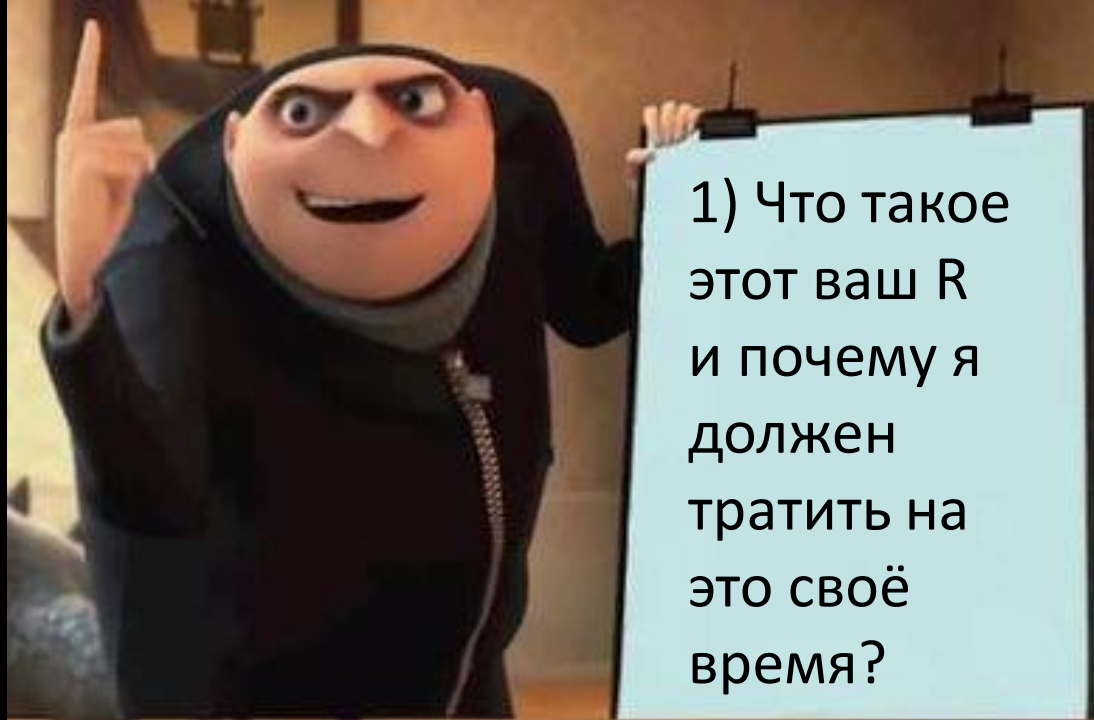
к.б.н., н.с. лаб.

геоинформационных технологий

ИЭРиЖ УрО РАН

[A.N.Sozontov | α | gmail | . | com](mailto:A.N.Sozontov@gmail.com)

ipae.uran.ru/Sozontov_AN



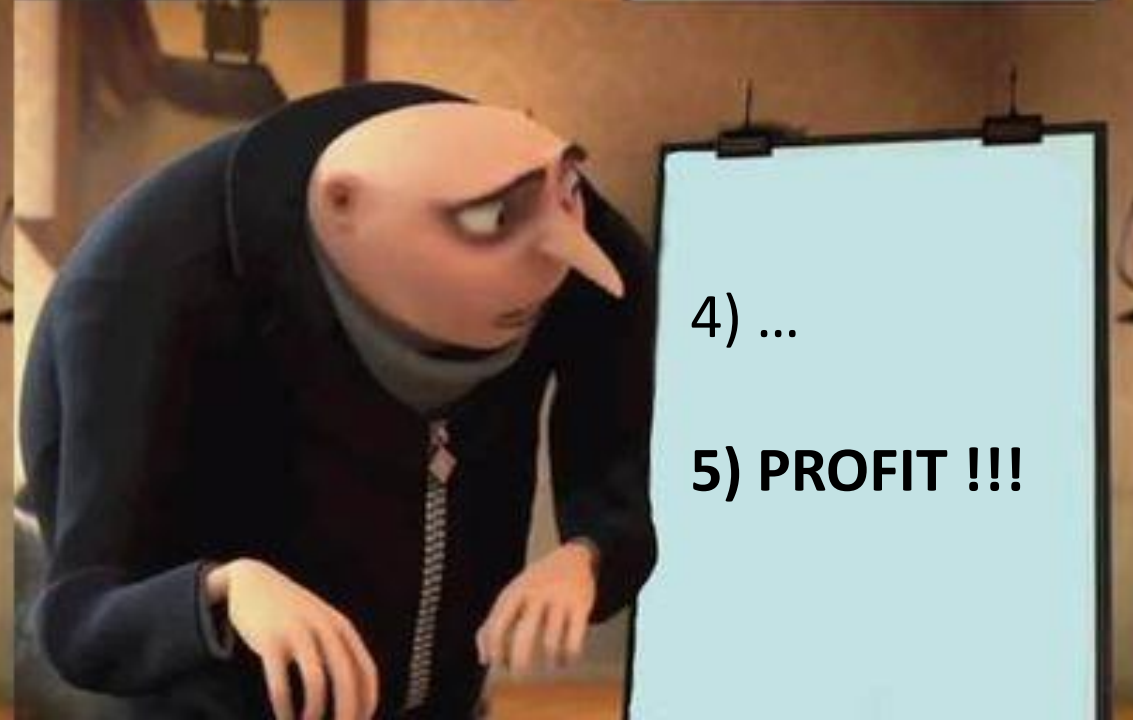
1) Что такое
этот ваш R
и почему я
должен
тратить на
это своё
время?



2) Чего R
умеет такого,
что я не могу
сделать в
Excel / Past /
Statistica и
другом ПО?



3) Считать это
ладно, но
сумеет ли R в
два клика
вывести мне
на экран
график, карту,
диплом или
диссертацию?



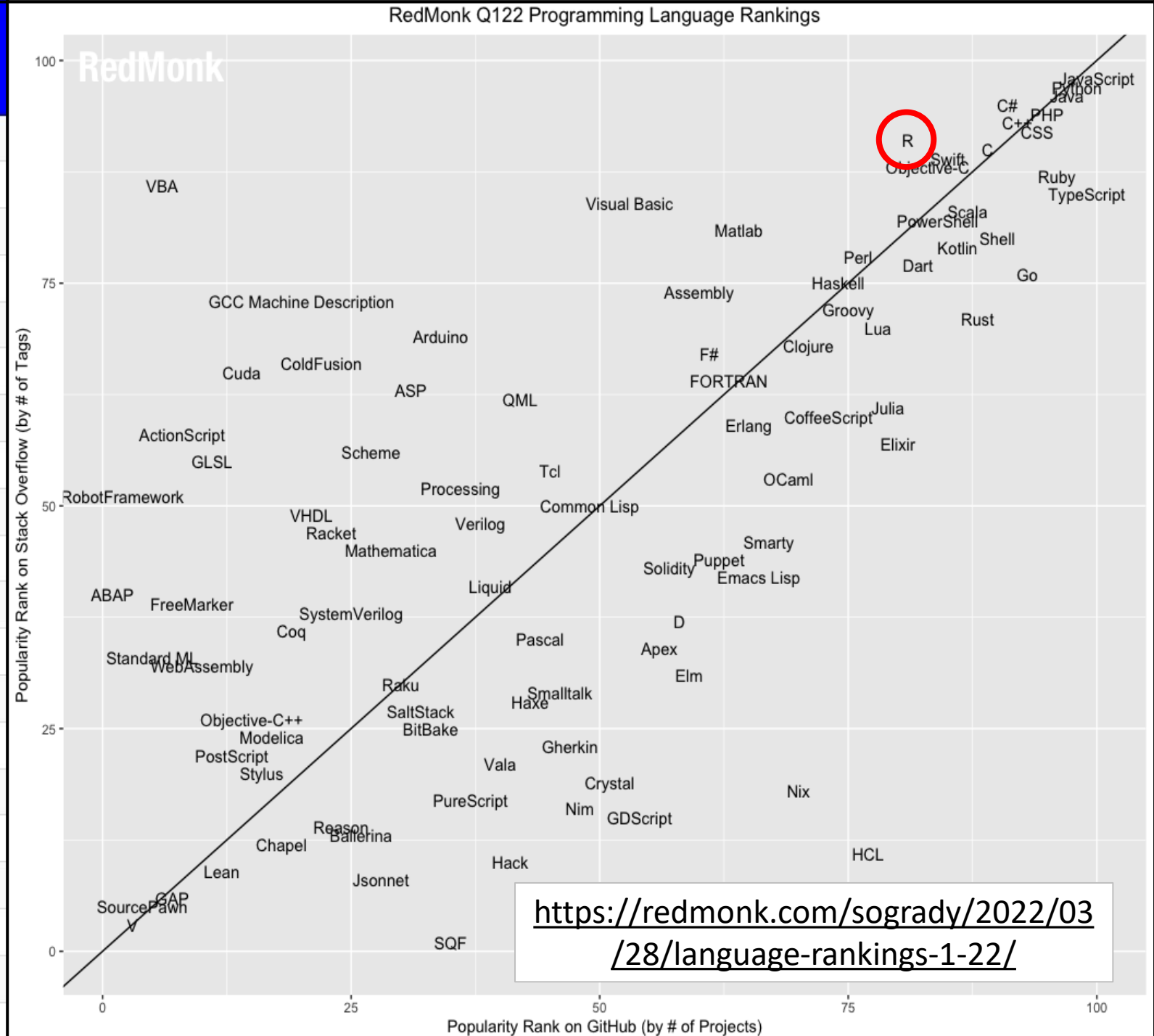
4) ...

5) PROFIT !!!

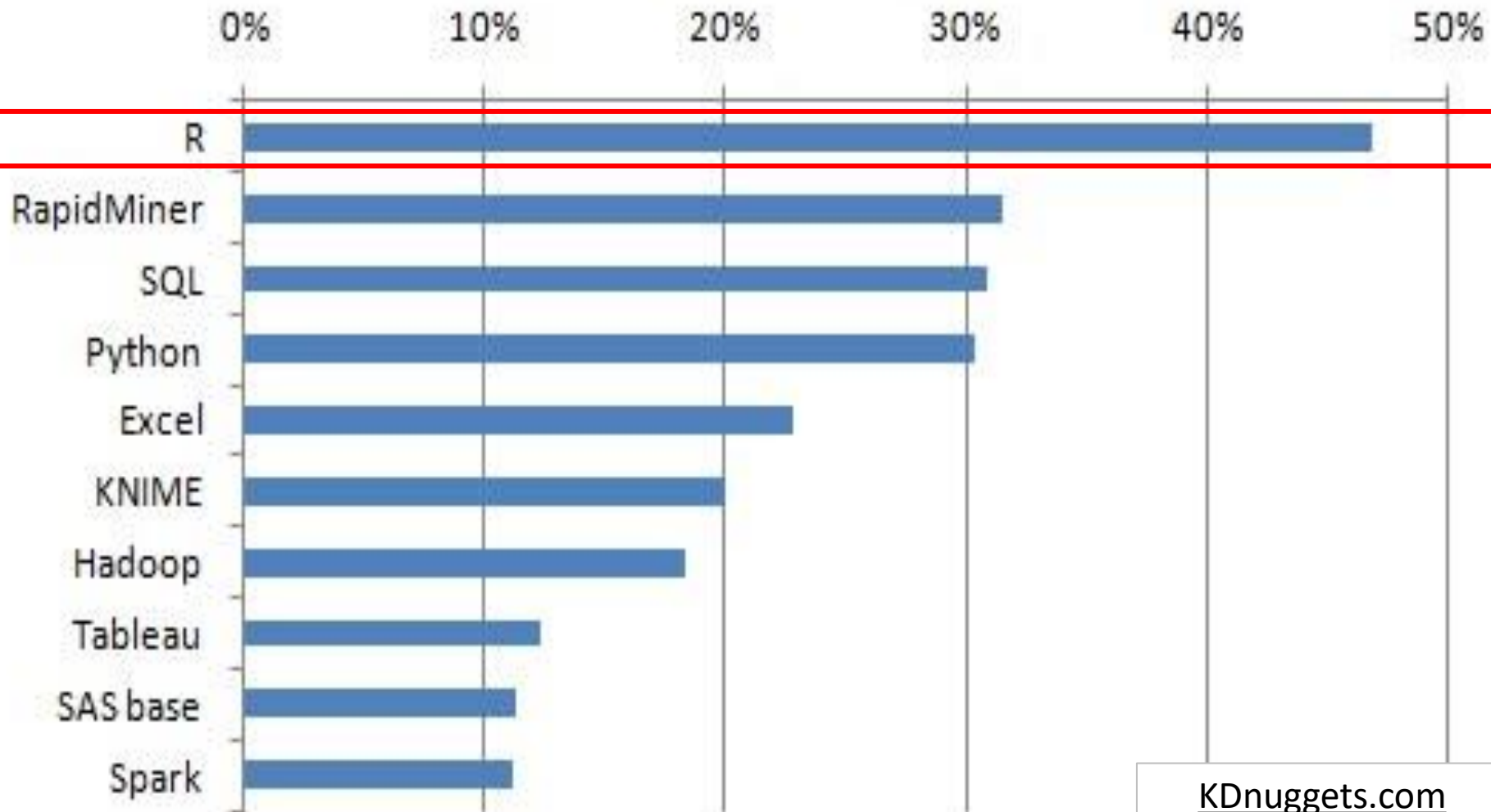
Position	Language	release	Designed by	Influenced by
#1	Python	Python 3.10.4	Guido van Rossum	ABC, Ada, ALGOL 68, APL, C, C++, CLU, Dylan, Haskell, Icon, Java, Lips, Modula-3, Perl, Standard ML
#2	Java	Java SE 17	James Gosling	Ada 83, C#, C++, CLU, Eiffel, Lisp, Mesa, Modula-3, Oberon, Object Pascal, Objective-C, Smalltalk,
#3	JavaScript	ECMAScript	Brendan Eich	AWK, HyperTalk, Java, Scheme
#4	C# (C Sharp)	C# 10.0	Anders Hejlsberg	C++, Cω, Eiffel, F#, Haskell, Icon, J#, J++, Java, ML, Modula-3, Object Pascal, VB
#5	PHP	PHP 8.1.4	Rasmus Lerdorf	C, C++, Hack, HTML, Java, JavaScript, Perl, Tcl
#6	C	C17	Dennis Ritchie	ALGOL 68, Assembly, B (BCPL, CPL), FORTRAN,
#7	R	R 4.1.3	Ross Ihaka and Robert Gentleman	Common Lisp, S, Scheme, XLispStat
#8	TypeScript	TypeScript	Microsoft	C#, Java, JavaScript
#9	Swift	Swift 5.6	Chris Lattner, Doug Gregor, John	C#, CLU, D, Haskell, Objective-C, Python, Ruby, Rust
#10	Objective-C	Objective-C 2.0	Tom Loe and Brad Cox	C, Smalltalk

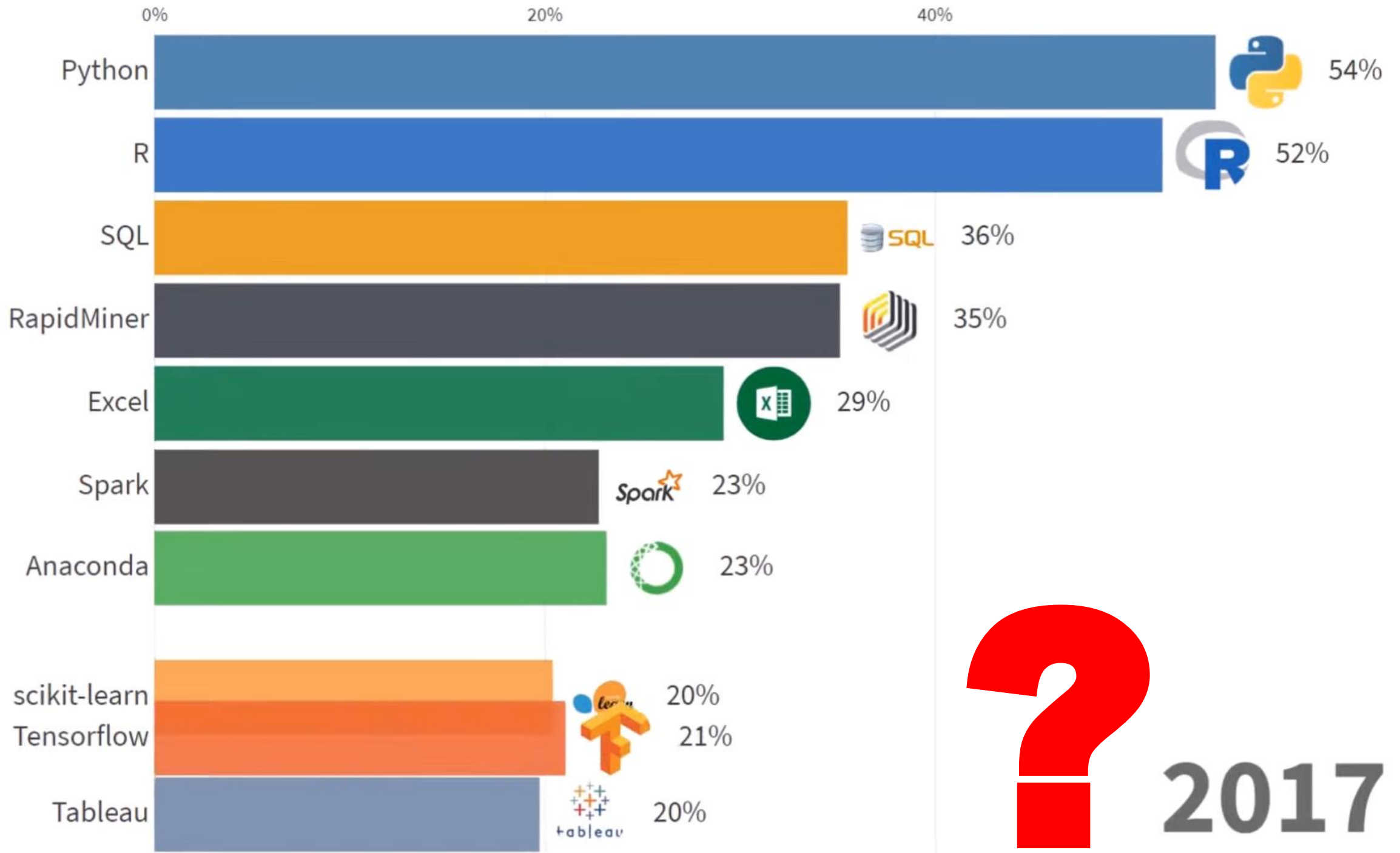
Aug 2020	Change	Programming language	Ratings
1		C	12.57%
3	↑	Python	11.86%
2	↓	Java	10.43%
4		C++	7.36%
5		C#	5.14%
6		Visual Basic	4.67%
7		JavaScript	2.95%
9	↑	PHP	2.19%
14	↑↑	Assembly language	2.03%
10		SQL	1.47%
18	↑↑	Groovy	1.36%
17	↑↑	Classic Visual Basic	1.23%
42	↑↑	Fortran	1.14%
8	↓↓	R	1.05%
15		Ruby	1.01%
12	↓↓	Swift	0.98%
16	↓	MATLAB	0.98%
11	↓↓	Go	0.90%
36	↑↑	Prolog	
13	↓↓	Perl	0.78%

tiobe.com



Top Analytics, Data Mining, Data Science software used, 2015





[←](#) Tweet

Dirk Eddebuettel
@eddelbuettel



[#ThankYouCRAN](#), and congratulations on another round number of [#RStats](#) packages -- now at 18,000. Just wow.

Available Packages

Currently, the CRAN package repository features 18000 available packages.

4:55 PM · Aug 11, 2021 · Twitter Web App

19 Retweets 2 Quote Tweets 126 Likes



 Tweet your reply

Reply

 Search Twitter

Relevant people



Dirk Eddebuettel
@eddelbuettel

Follow

Data Science. TileDB. Open Source.
Quant Research. R. C++. Debian.
Linux. Adjunct Clinical Professor,
University of Illinois. Lots of coffee.
And some running.

Trends



1 · Trending



[#russianmcyttwtselfieday](#)

1,113 Tweets

2 · Space · Trending



[best news of the week](#)

3 · Trending



R Package Documentation

A comprehensive index of R packages and documentation from CRAN, Bioconductor, GitHub and R-Forge.

Search for anything R related

Find an R package by name, find package documentation, find R documentation, find R functions, search R source code...

gaussian

Search

21932

CRAN PACKAGES

2130

BIOCONDUCTOR PACKAGES

2199

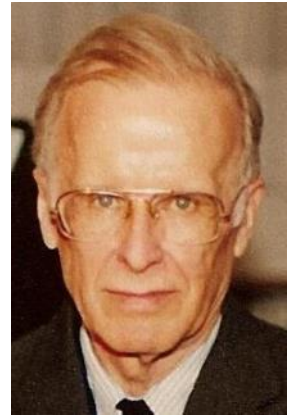
R-FORGE PACKAGES

84436

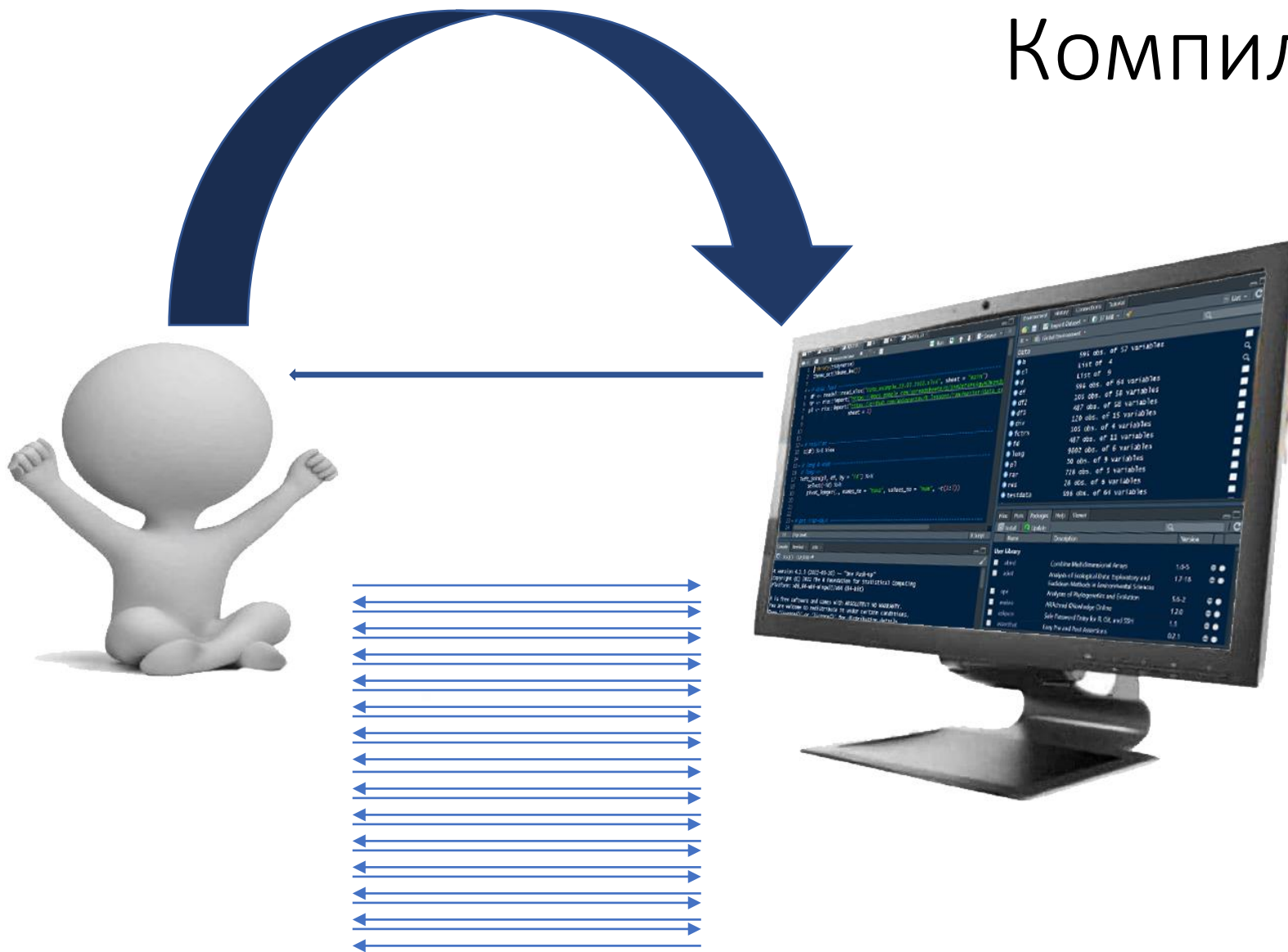
GITHUB PACKAGES

Ключевые события

- 1954 – John Backus в IBM разрабатывает первый язык программирования высокого уровня FORTRAN



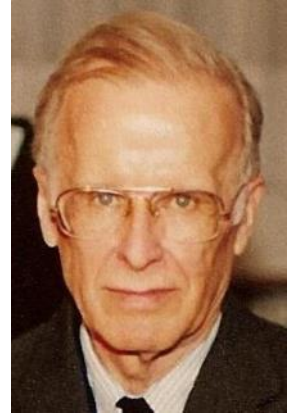
Компилируемый



Интерпретируемый

Ключевые события

- 1954 – John Backus в IBM разрабатывает первый язык программирования высокого уровня FORTRAN



- 1976 – John Chambers (Bell Labs) начинает разработку языка S, способного *«превратить идеи в программное обеспечение быстро и точно»*
- 1988 – S-PLUS

- 1993 – Robert Gentleman и Ross Ihaka начали разрабатывать R как бесплатную альтернативу среде S



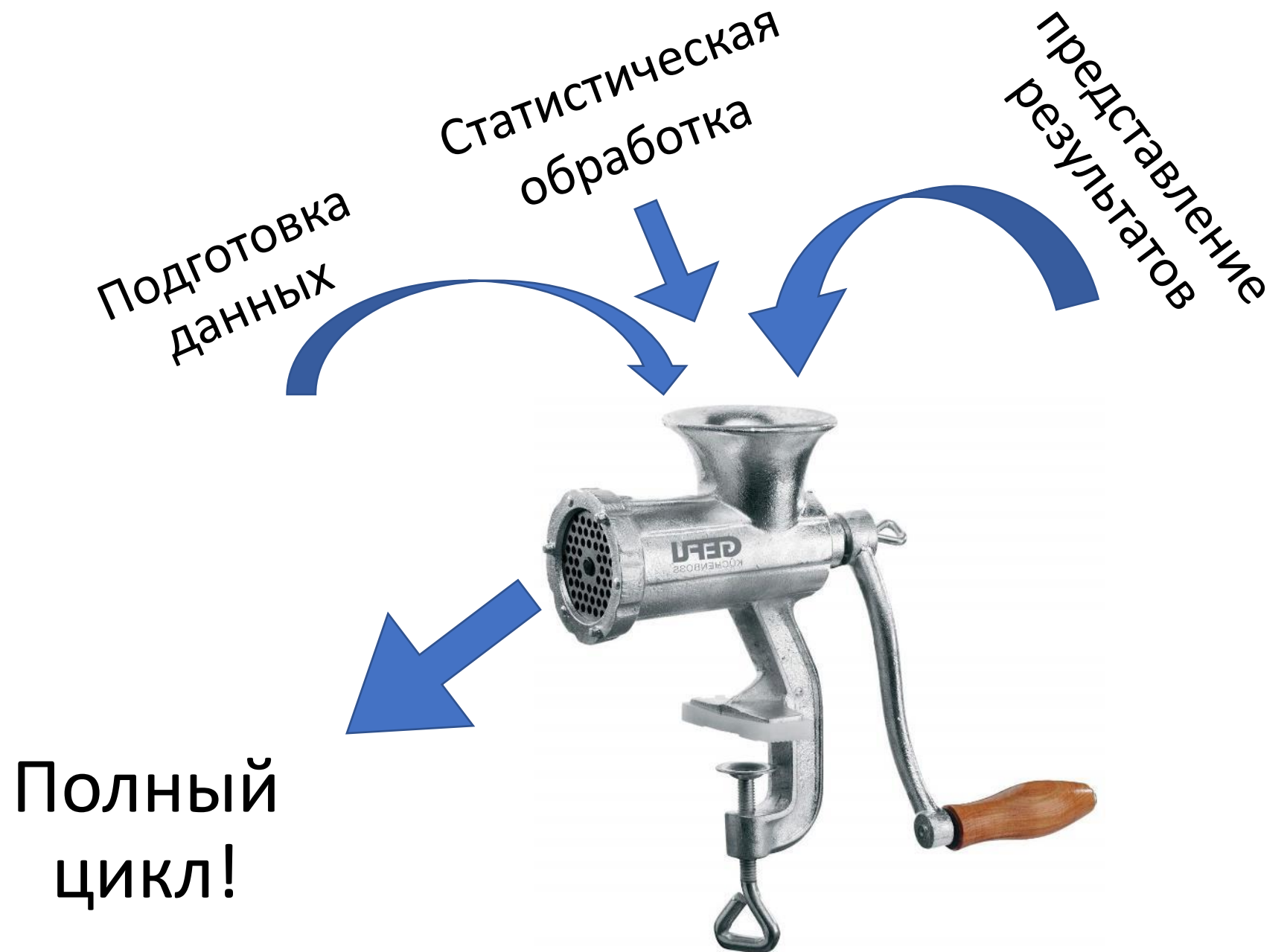
- 2005 – Hadley Wickham создал ggplot2
- 2011 – RStudio
- 2014 – Hadley Wickham выпустил пакет dplyr v.0.1
- 2022: R, v.4.1.3

Лицензия


R распространяется под лицензией

GNU GPL (General Public License), что означает:

- Каждый может использовать R без ограничений
- Каждый может модифицировать R в своих целях без ограничений
- Каждый может распространять R (в том числе и свою, улучшенную версию) без ограничений

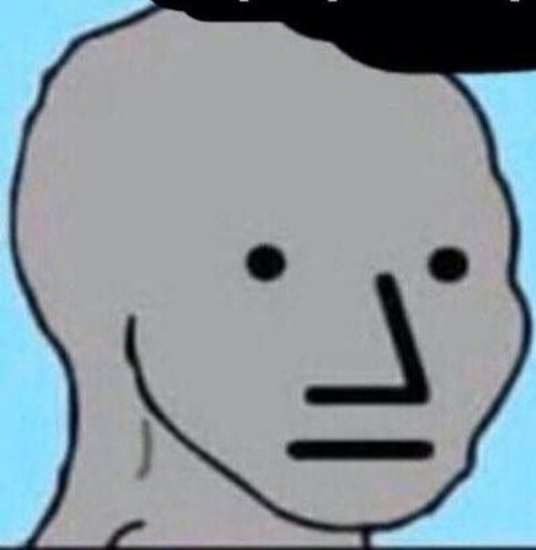


1.  =   
StatSoft®

2.  > >   
StatSoft®

3.  !=   
StatSoft®

Хочу научиться
программировать



Тогда научись



@cantbeparsedinreasonabletime





R Console

```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R -- это свободное ПО, и оно поставляется безо всяких гарантий.
Вы вольны распространять его при соблюдении некоторых условий.
Введите 'license()' для получения более подробной информации.

R -- это проект, в котором сотрудничает множество разработчиков.
Введите 'contributors()' для получения дополнительной информации и
'citation()' для ознакомления с правилами упоминания R и его пакетов
в публикациях.

Введите 'demo()' для запуска демонстрационных программ, 'help()' -- для
получения справки, 'help.start()' -- для доступа к справке через браузер.
Введите 'q()', чтобы выйти из R.

[Загружено ранее сохраненное рабочее пространство]
```

```
> |
```


Минус Решение

Нет русскоязычного интерфейса и справки

- Набор команд ограничен, есть руководства на русском языке

Интерфейс R не дружелюбный

- Существуют оболочки для R, предоставляющие оконный интерфейс
- R можно интегрировать в Statistica (>10), SPSS и MS Excel

Чтобы работать в R, нужно уметь программировать

- Да, но научиться основам достаточно просто; уже базовые навыки позволяют автоматизировать однотипные операции и тем самым ускорять решение многих задач (вся мощь R кроется именно в возможности писать скрипты)

Чувствителен к синтаксису

- Смириться и быть внимательнее

Не существует технической поддержки R

- Существует обширная справочная и учебная литература по работе в R (в т.ч. на русском), большое сообщество пользователей, многие из которых готовы помочь.

Интерпретируемый и потому медленный

- Рядовой пользователь не заметит, что скорость вычислений низкая в сравнении с компилируемыми или JIT-компилируемыми языками программирования
- Имеются простые и эффективные средства распараллеливания вычислений
- Предоставляет базу для дальнейшего освоения более сложных и производительных языков (Python, Julia, Go, Ruby)

Сложно ориентироваться в многообразии пакетов и функций

- Использовать хорошие практики, узнать о которых можно из руководств, справки, форумов, обучающих видео и, конечно же, от своих коллег

Плюс Применение

Свободная лицензия	• Свобода и бесплатность использования
Автоматизация	• Экономия времени за счет автоматизации рутинных операций
Концепция «Все в 1»	• Не нужно учить много программ, достаточно одной для решения всех задач
Предобработка	• Простота и скорость чистки, доработки и преобразования первичных данных
Статистика	• Самый полный набор всевозможных статистических процедур: поправки на множественное сравнение, GLM с любыми распределениями, всевозможные варианты бутстрепа, весь арсенал методов многомерной статистики и т.д.
Графика	
Распараллеливание	• Ресурсоемкие вычисления можно делать эффективнее
ГИС-технологии	• Скорости обработки карт и других пространственных данных
Отчеты RMarkdown	• Не надо переделывать всю диссертацию, если есть изменения в данных
Веб-приложения Shiny	• Зачем запускать код многократно, если можно один раз и пойти пить чай?
SDM и rgbif	• Прямой доступ к данным с GBIF и использование их для моделирования ареалов

И все это только вершина айсберга...

Протоколирование	• Остаются задокументированы <u>ВСЕ</u> этапы работы с данными. Это сочетается с концепциями FAIR-data и OpenScience, улучшает воспроизводимость, позволяет коллегам учиться или находить ошибки.
------------------	---

УДК 574.3:599.32/.38:502.175:[504.5:669.2/.8](470.54)

МНОГОЛЕТНЯЯ ДИНАМИКА СООБЩЕСТВ МЕЛКИХ МЛЕКОПИТАЮЩИХ В ПЕРИОД СНИЖЕНИЯ ВЫБРОСОВ МЕДЕПЛАВИЛЬНОГО ЗАВОДА. II. БЕТА-РАЗНООБРАЗИЕ

В статистических тестах значимыми считали различия при $p < 0.05$. Бутстрепные доверительные интервалы для β_W и I_{BC} получены на основе 9999 итераций. Расчеты и визуализация выполнены в среде программирования R v.4.0.3 [35] при помощи пакетов *ape* [36], *vegan* [37] и *ggplot2* [38]. Исходные данные и код размещены по адресу: github.com/ANSozontov/betadiv_2020.

УДК 569.72/.73:574.34(470-924.85+571-925.116)"627"

ИЗМЕНЕНИЯ СОСТАВА И ОТНОСИТЕЛЬНОГО ОБИЛИЯ КОПЫТНЫХ ЕВРАЗИЙСКОЙ ЛЕСОСТЕПНОЙ ЗОНЫ В ГОЛОЦЕНЕ

Расчеты и визуализация выполнены в среде программирования R 4.02 (R Core Team, 2020) при помощи пакетов *ape* (Paradis, Schliep, 2019) и *vegan* (Oksanen et al., 2019) для обработки многомерных данных (коэффициенты сходства, PCoA, PERMANOVA), а также пакета *ggplot2* (Wickham, 2016) и коллекции пакетов *tidyverse* (Wickham et al., 2019) для визуализации и предварительной обработки данных. Все скрипты и исходные данные доступны на репозитории GitHub по адресу: https://github.com/ANSozontov/Gasilin_2020.

Package 'adegenet'

October 9, 2021

Title Exploratory Analysis of Genetic and Genomic Data

Version 2.1.5

Description Toolset for the exploration of genetic and genomic data. Adegenet provides formal (S4) classes for storing and handling various genetic data, including genetic markers with varying ploidy and hierarchical population structure ('genind' class), alleles counts by populations ('genpop'), and genome-wide SNP data ('genlight'). It also implements original multivariate methods (DAPC, sPCA), graphics, statistical tests, simulation tools, distance and similarity measures, and several spatial methods. A range of both empirical and simulated datasets is also provided to illustrate various methods.

License GPL (>= 2)

URL <https://github.com/thibautjombart/adegenet>

Depends R (>= 2.14), methods, ade4

Imports utils, stats, grDevices, MASS, igraph, ape, shiny, ggplot2, seqinr, parallel, boot, reshape2, dplyr (>= 0.4.1), vegan

Suggests adespatial, pegas, hierfstat, akima, maps, spdep, splancs, tripack, testthat, poppr

Encoding UTF-8

NeedsCompilation yes

Author Thibaut Jombart [aut] (<<https://orcid.org/0000-0003-2226-8692>>),
Zhian N. Kamvar [aut, cre] (<<https://orcid.org/0000-0003-1458-7108>>),
Caitlin Collins [ctb],
Roman Lustrik [ctb],
Marie-Pauline Beugin [ctb],
Brian J. Knaus [ctb],
Peter Solymos [ctb],
Vladimir Mikryukov [ctb],
Klaus Schliep [ctb],
Tiago Maié [ctb],
Libor Morkovsky [ctb],
Ismail Ahmed [ctb],
Anne Cori [ctb],
Federico Calboli [ctb],
RJ Ewing [ctb],
Frédéric Michaud [ctb],
Rebecca DeCamp [ctb],
Alexandre Courtiol [ctb] (<<https://orcid.org/0000-0003-0637-2959>>)

Maintainer Zhian N. Kamvar <zkamvar@gmail.com>

Repository CRAN

Date/Publication 2021-10-09 18:20:01 UTC

R topics documented:

Рекомендуемая литература на русском

- **Кабаков Р.И. (2014)** R в действии. Анализ и визуализация данных в программе R. М.: ДМК Пресс. 588 с.
- **Шипунов А.Б. и др. (2014)** Наглядная статистика. Используем R! [Электронная книга]
- **Мастицкий С.Э., Шитиков В.К. (2014)** Статистический анализ и визуализация данных с помощью R [Электронная книга]
- **Шитиков В.К., Мастицкий С.Э. (2017)** Классификация, регрессия и другие алгоритмы Data Mining с использованием R [Электронная книга]
- **Эрве М. (2016)** Путеводитель по применению статистических методов с использованием R. Планирование исследований и анализ результатов в биологии с помощью программного обеспечения R [Электронная книга]
- **Зарядов И.С. (2010)** Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Изд-во РУДН. 207 с.
- **Зарядов И.С. (2010)** Статистический пакет R: теория вероятностей и математическая статистика. М.: Изд-во РУДН. 141 с.

Спасибо за внимание

Выражаю признательность

- **Самсонову Тимофею** за пол года критических комментариев моего кода
- **Микрюкову Владимиру** за серию консультаций
- **Модорову Макару** за 20 минутный ликбез по R, вдохновивший меня на дальнейшее самостоятельное изучение
- **Трубицыну Андрею (1978–2009)** за первое (но чрезвычайно воодушевляющее) знакомство с миром электронных вычислительных машин

Важные ссылки

- Папка на гитхабе для материалов:
https://github.com/ANSozontov/R_lessons
- Ядро R: cran.r-project.org/bin/windows/base
- RTools: cran.r-project.org/bin/windows/Rtools
- Rstudio: www.rstudio.com/products/rstudio/download/#download
- Самосонов Т. «Визуализация и анализ географических данных на языке R»: tsamsonov.github.io/r-geo-course
- Курс «Основы программирования на R»: stepik.org/course/497

