Statistical programming languages

Introduction to Statistical Programming

Introduction to Statistical Programming

The purpose of the lecture is to orient students in the field of technologies and methodologies for analyzing big data, to gain knowledge about the main tasks facing the science of data, about the software used in this area.

As a result of studying the lecture materials, you will know what data science is, what skills a specialist in this field should have, what software tools help to analyze big data.

Since 2013 BIG DATA as an academic subject is studied in the emerging university programs on the subject DATA SCIENCE

wikipedia.org

Lecture questions:

1. The purpose and content of the course
2. What is a Data Science, who is a Data Scientist and what should he be able to do?
3. Big data exploration software
4. Areas of application and examples of using the programming languages R and Python

Literary sources:

1. Data Science Skills. Alexey Voronin. Source: https://habrahabr.ru/post/271085/

2. Do you need to learn the R language? Katherine Delzell. Source: https://www.ibm.com/developerworks/ru/library/bd-learnr/

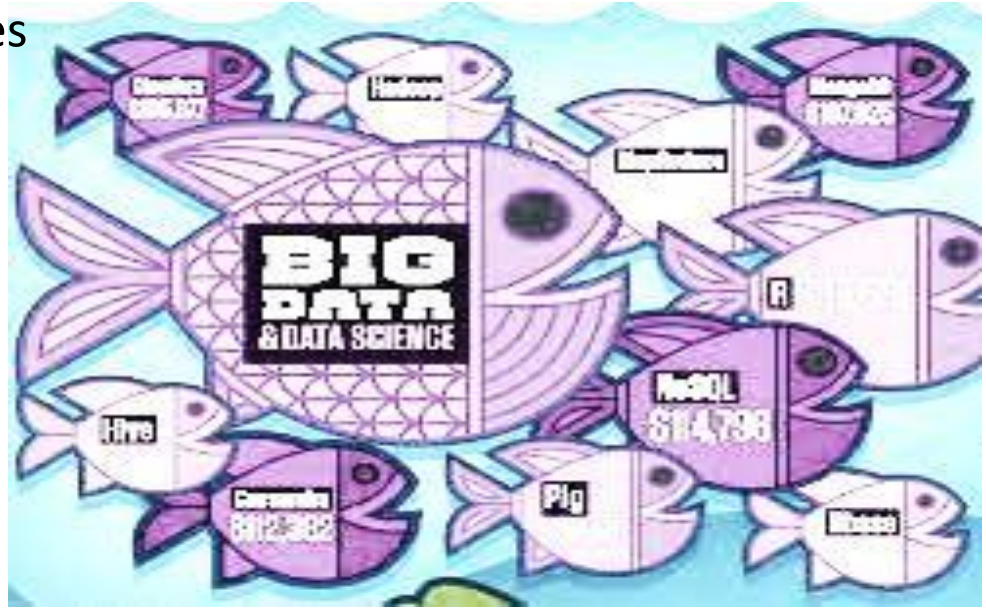
3. Python 3 programming language for beginners and dummies. Portal: https://pythonworld.ru/

Data

Is an ocean full of sea creatures

but until they are caught,

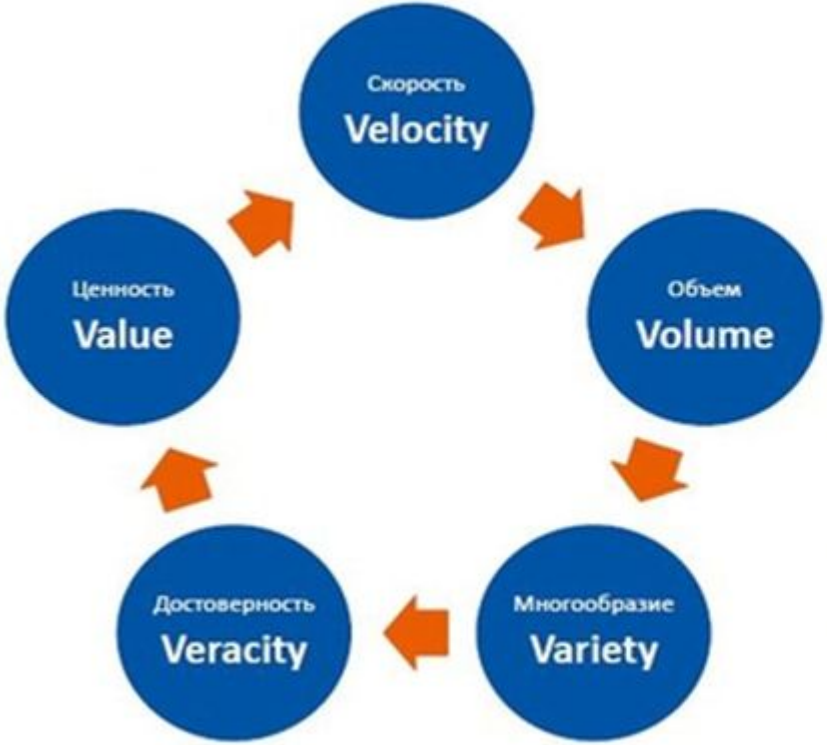
no benefit from them !!!



Differences between traditional databases and Big Data

Characteristic	Traditional database	Big data database
Amount of information	From gigabytes (10^9 bytes) to terabytes (10^12 bytes)	From petabytes (10^15 bytes) to exabytes (10^18 bytes)
Storage method	Centralized	Decentralized
Data structuring	Structured	Semi-structured or unstructured
Data storage and processing model	Vertical model	Horizontal model
The relationship of data	strong	weak

Differences between traditional databases and Big Data



Characteristic	Traditional database	Big data database
Amount of information	From gigabytes (10^9 bytes) to terabytes (10^12 bytes)	From petabytes (10^15 bytes) to exabytes (10^18 bytes)
Storage method Data structuring	Centralized Structured	Decentralized Semi-structured or unstructured
Data storage and processing model The relationship of data	Vertical model strong	Horizontal model weak

http://www.tadviser.ru/index.php

Global data growth

90% of all information generated over the past 2 years

— SINTEF

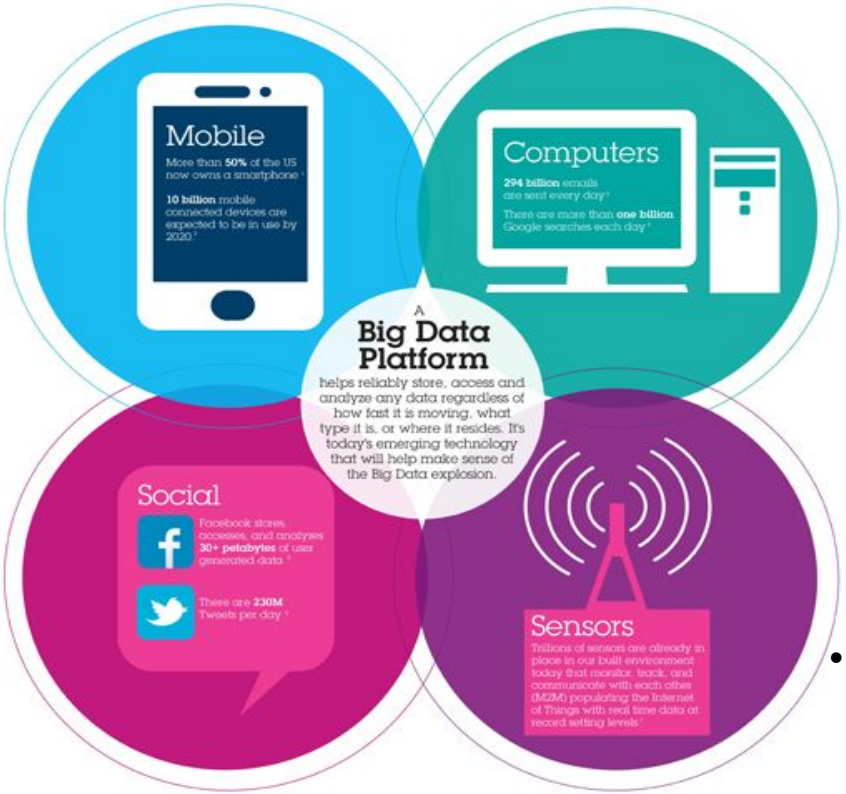
Facebook stores and processes over 50 Tb

Twitter generates per day 8 Tb

10 trillion gigabyte annual amount of data processed in 2016

— University of California

Big Data Sources



1. Social Networks

2. Machine data

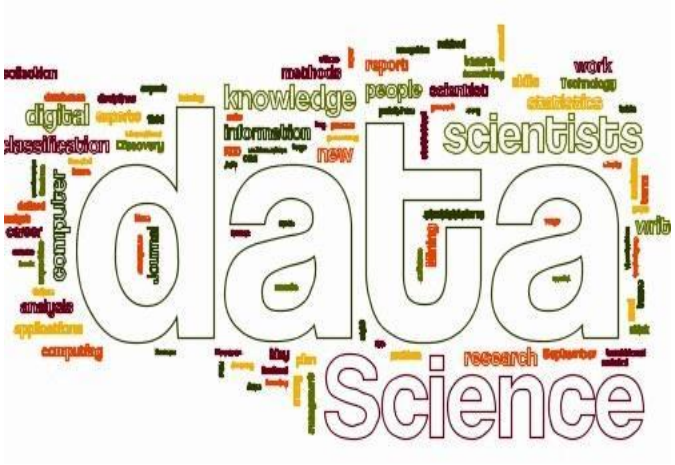
3. Transaction Data

They can also be divided into:

- current and historical obtained from open and closed sources, structured and unstructured.

Data science is a new discipline that draws on knowledge in statistical methodology and computer science to create impressive forecasts and ideas for a wide range of traditional scientific fields.http://datascience.harvard.edu/

Directions of research in the field of Data Science



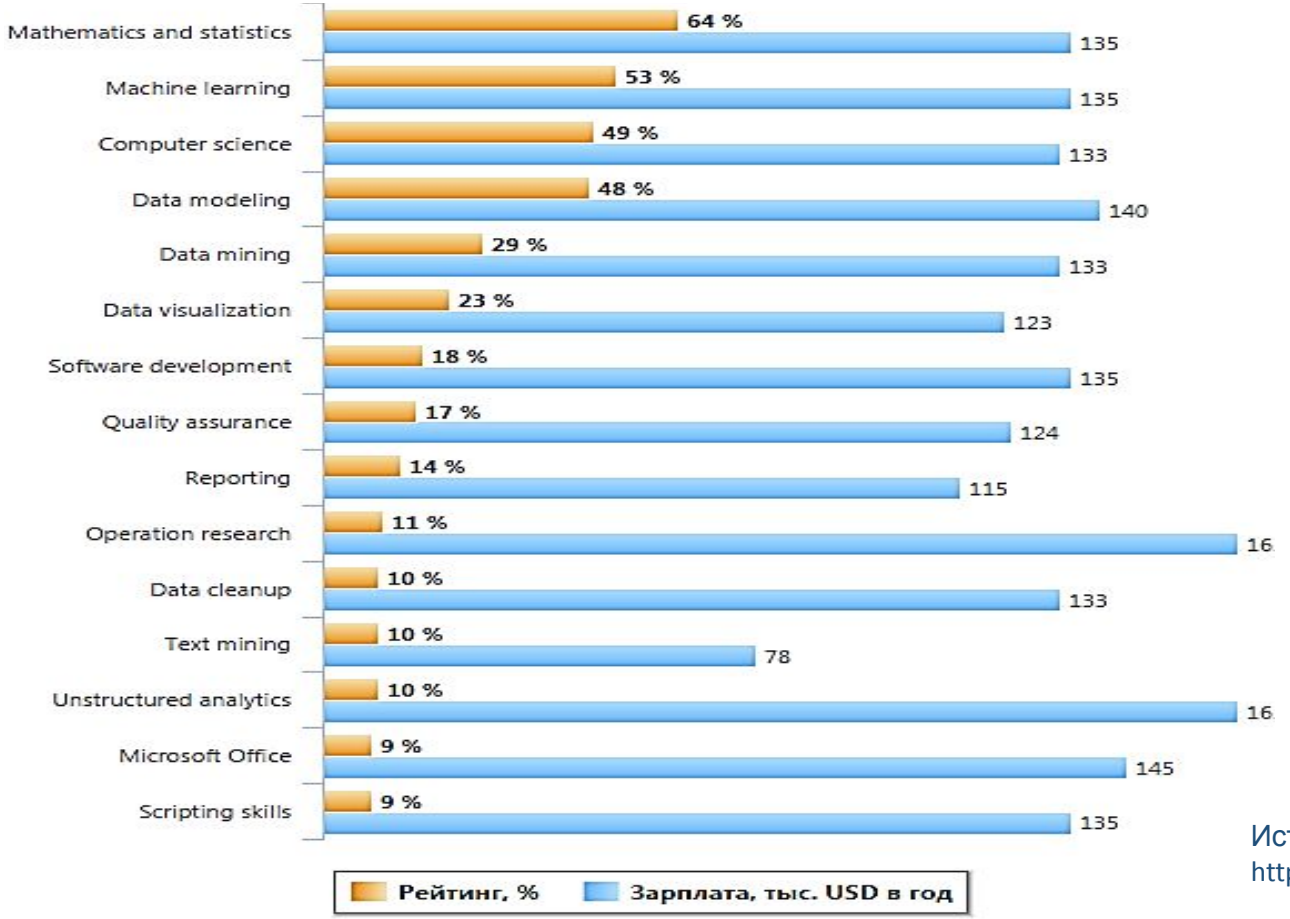
- Cloud computing
- Databases and information
- integration Signal processing
- Learning,
- Natural Language Processing, and Information Retrieval
- Computer vision
- Information Search
- Discovery of knowledge in social and information networks
- Information visualization

Data Scientist - data scientist is a kind of hybrid statistics and programmer

- this is someone who understands statistics better than any programmer,

and better versed in programming than any statistician.

Proficiency Requirements **(hard skills)**



Источник:
https://habrahabr.ru/post/271085/

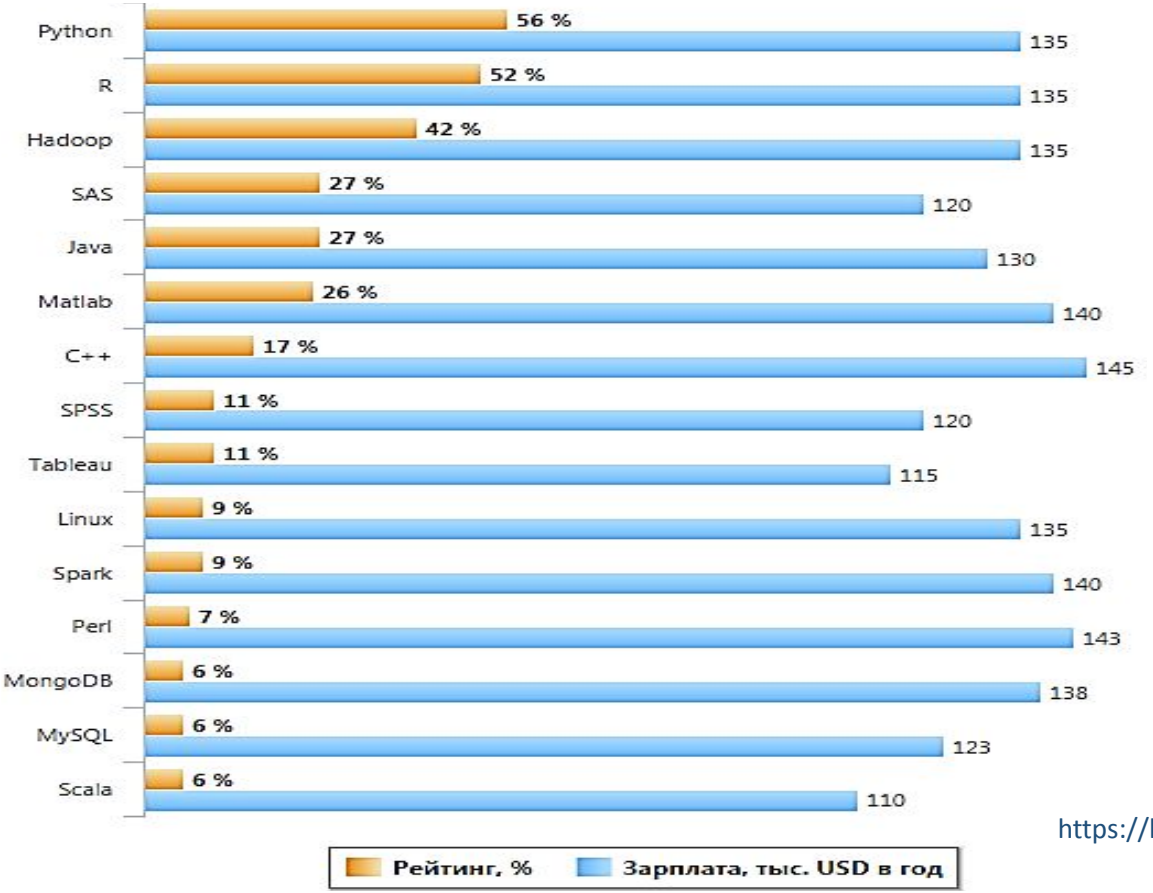
Statistical programming languages

What is advisable to know before learning the R and Python languages?**?**

- Statistical Data Analysis Methods
- Probability theory
- Mathematical analysis
- Linear algebra
- Data mining

Wikipedia tells us that to date, dozens of software products have already been developed for data analysis, in particular, statistical processing. Consider briefly the most popular among them.

The core Data Scientist toolkit is the Python and R programming languages



https://habrahabr.ru/post/271085/

Statistical programming languages

Statistical tools can be divided into three types :

- programs with a graphical interface based on the principle of “click here and get the finished result” (PRISM, Statex);
- statistical programming languages that require basic R and Python programming skills;
- "mixed", in which there is a graphical interface (GUI), and the ability to create script programs (for example: SAS, STATA, Rcmdr).

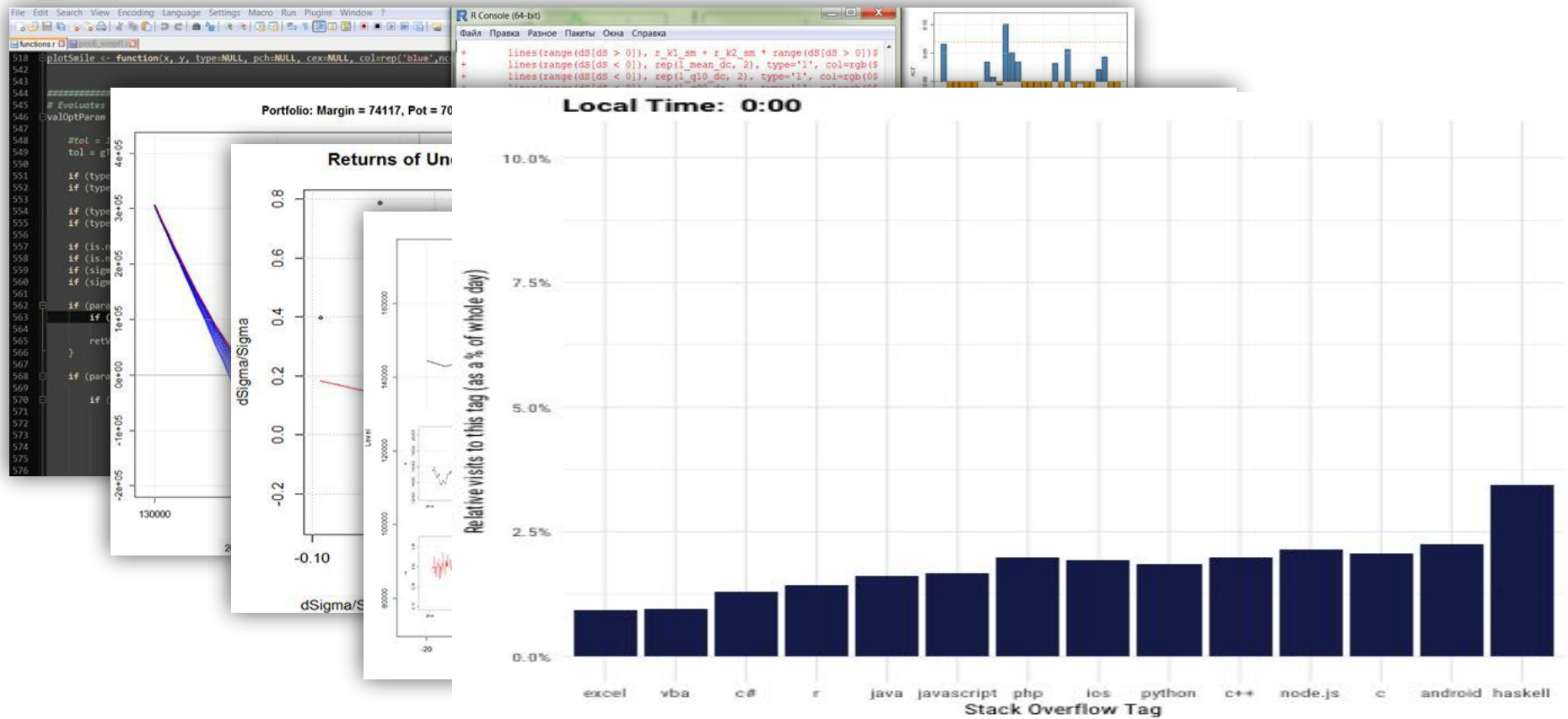
What is R?

1. Programming language and development environment for statistical computing and graphics GNU Open Source Project
2. A variety of statistical and graphical methods (linear and non-linear modeling, statistical analysis, time series analysis, cluster analysis, ...)
3. Functionality greatly expanded with packages
4. Works under UNIX, Windows, MacOShttp://www.r-project.org/

Why R?

- Absolutely free
- A language specifically designed for statistical analysis
- Huge data visualization capabilities
- Over 5000 extension packs
- Develops faster than any commercial software
- Hundreds of books, “The R Journal”, “Journal of Statistical Software”
- A huge number of users (> 3 million, 2016)
- Support, fast error correction

R graphics capabilities



HISTORY OF THE R LANGUAGE

R -dialect of SqlS was created in 1976 at Bell Labs

"R is a programming language for statistical data processing and graphics, as well as a free and open source computing environment under the GNU project.»

Wikipedia

The R language was created in 1991 by statisticians Ross And Haka and Robert Gentleman (University of Auckland, New Zealand)

2. Installation

R:

http://cran.r-project.org

RStudio:

RStudio: http://rstudio.org

2. R

Надежный | https://cran.r-project.org

Сервисы | Создание лаборатор | Введение в R: часть | Введение в R: часть | Введение в R: часть | Введение в R: часть | Введение в R: часть | Из Google Chrome

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Friday 2017-06-30, Single Candle) R-3.4.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

Installation file

Надежный | https://cran.r-project.org

Сервисы | Создание лаборатор | Введение в R: часть | Введение в R: часть | Введение в R: часть | Введение в R: часть | Введение в R: часть | Из Google Chrome

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

R for Windows

Subdirectories:

base	Binaries for base distribution (managed by Duncan Murdoch). This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.11.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.11.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

www.cran.r-project.rg - 146 тыс. | The Comprehensive R Archive Ne

cran.r-project.org

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2019-07-05, Action of the Toes) R-3.6.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

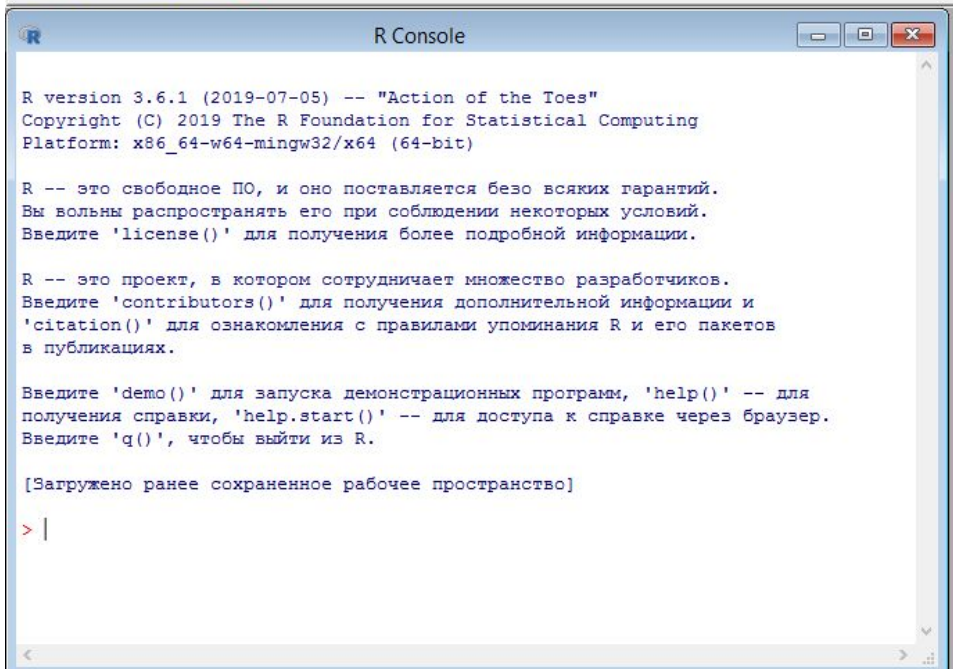
Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

3 RGUI

RGui is the standard that comes with the package itself. RGui is fast to download and quite easy to use.
It has three kinds of windows:
console;
the script window;
graphics device window.
In the console, R commands are typed and sent to execute (by pressing Enter)

3 R GUI


R Console
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 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]
>

4. RStudio: http://rstudio.org

RStudio

File Edit View Workspace Plots Help

diamondPricing.R* | diamonds

Source on Save | Run Line(s) | Run All

```
library(ggplot2)

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat),4)
clarity <- levels(diamonds$clarity)

qplot(price, carat, data = diamonds)

qplot(price, carat, data = diamonds, color=clarity,
    xlab = "Price", ylab = "Carat",
    main = "Diamond Pricing") +
    opts(plot.title = theme_text(size = 22))
```

Workspace | History

Load | Save | Import Dataset | Clear All

Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]

Files | Plots | Packages | Help

Zoom | Export | Print | Clear All

Console ~/

```
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800

> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat),4)
> clarity <- levels(diamonds$clarity)
> qplot(price, carat, data = diamonds)
> qplot(price, carat, data = diamonds, color=clarity, xlab =
"Price", ylab = "Carat", main = "Diamond Pricing") +
opts(plot.title = theme_text(size = 22))
>
```

Integrated development environment (IDE) for R

Combines an intuitive interface with powerful R code development tools

R Studio is an integrated development environment (IDE)

script window

console window

workspace,

command history

working folder,

graphics, installedpackages

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

script.R*

Source on Save

Run

Source

```
library(datasets)

# Рассмотрим таблицу с данными о скорости
# движения машин и их тормозного пути (cars)

str(cars)
summary(cars)
head(cars)

# Визуализируем данные и корреляцию между скоростью и
# тормозным путем
attach(cars)
hist(speed, breaks = 10, col = "light blue")
hist(dist, breaks = 10, col = "pink")
plot(speed, dist, main = "Corr: speed and dist", pch=21,
     bg="lightgreen")
cor(speed, dist)
legend("topleft", "R = 0.81")
detach(cars)
```

15:1 (Top Level) R Script

Console ~/

```
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
> # Визуализируем данные и корреляцию между скоростью и
> # тормозным путем
> attach(cars)
> hist(speed, breaks = 10, col = "light blue")
>
```

Environment History

Import Dataset List

package:datasets

Data

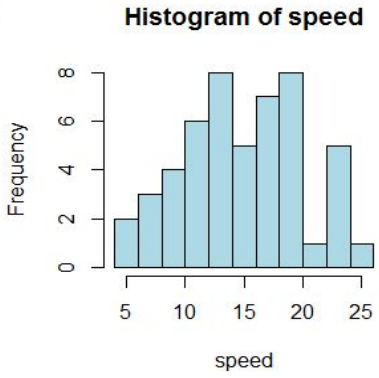
cars 50 obs. of 2 variables

values

ability.... <Promise>
airmiles <Promise>
AirPasse... <Promise>
airquality <Promise>
anscombe <Promise>

Files Plots Packages Help Viewer

Zoom Export



4. RStudio: installation file

iGoogle | R: Анализ и визуализация | RStudio

rstudio.org/download/

Home | Screenshots | Download | Docs | Support | Development | Blog

Download RStudio v0.94

If you run R on your desktop:

Download RStudio Desktop

OR

If you run R on a Linux server and want to enable users to remotely access RStudio using a web browser:

Download RStudio Server

RStudio - Google C... | Microsoft PowerPoi... | Total Commander 7... | RStudio | EN | 1:47 PM

4. RStudio: installation file

iGoogle | R: Анализ и визуализация | RStudio

rstudio.org/download/desktop

Home | Screenshots | Download | Docs | Support | Development | Blog

Download RStudio Desktop

RStudio v0.94 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it here.

Recommended For Your System	Size	Date	MD5
RStudio 0.94.84 - Windows XP/Vista/7	15.8 MB	2011-06-19	bebb9d530908f4561a7efe8cacf78fcb

All Platforms	Size	Date	MD5
RStudio 0.94.84 - Windows XP/Vista/7	15.8 MB	2011-06-19	bebb9d530908f4561a7efe8cacf78fcb
RStudio 0.94.84 - Mac OS X 10.5+	40.1 MB	2011-06-19	a7519d2c5b26ae8274a5854e29e5e184
RStudio 0.94.84 - Debian 6+/Ubuntu 10.04+ (32-bit)	23.3 MB	2011-06-19	7049a81e39a4a54649a82285f10b092b
RStudio 0.94.84 - Debian 6+/Ubuntu 10.04+ (64-bit)	23.7 MB	2011-06-19	a8ef4c1123c3b5f701131b5f8b899565
RStudio 0.94.84 - Fedora 13+ (32-bit)	23.3 MB	2011-06-19	dfc8fc26014ff4e59a6898f7fb5af476
RStudio 0.94.84 - Fedora 13+ (64-bit)	23.5 MB	2011-06-19	8c9abfbfa51139a65d50b47f60d1801d

Zip/Tarball

If you need an installer-less version of RStudio (for example, if you don't have administrative/root privileges on your computer) you can download a zip or tarball containing the RStudio binaries. Show zip/tarball downloads

Source Code

A tarball containing source code for RStudio v0.94.84 can be downloaded from here.

RStudio - Google C... | Microsoft PowerPoi... | Total Commander 7... | RStudio | EN | 1:48 PM

Exercises

- Go to the site R-project.org and check out its main sections
- From the “Documentation/Manuals” section, download the PDF files "An Introduction to R" and " R Data Import/Export”
- Note the “Documentation " section”

Introduction to Python

Python was created by Guido van Rossum in 1991.
Named the TV show after " Monty Python's flying circus»
The emphasis on performance and readability remains in this language.
Releases of the language:
Python 1.0-January 1994
Python 2.0-October 2000
Python 3.0-December 2008
Current versions:

- 2.7.8 Python
- Python 3.4.1

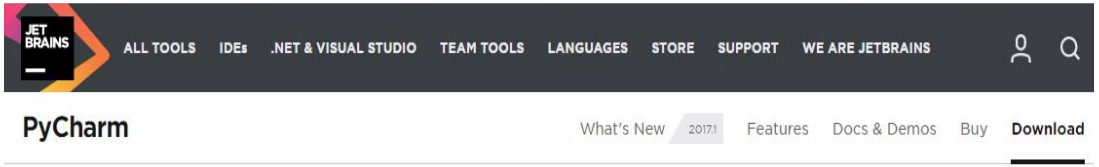
Advantages of using Python

- Software quality - Python code is easier to read, which means it is much easier to reuse and maintain
- Support libraries-Python allows expansion both through your own libraries and through libraries created by other developers
- Development speed - the amount of software code is usually a third, or even a fifth, of equivalent C++ or Java code
- Portability of programs to other operating platforms without changing the code

Software installation : Python 3.1

https://python.org/downloads/windows/

Software installation : PyCharm (IDE)



https://www.jetbrains.com/pycharm/download/

Download PyCharm

Windows macOS Linux

Version: 2017.1.1

Build: 171.4163.6

Released: April 12, 2017

System requirements

Professional

Full-featured IDE for Python & Web development

DOWNLOAD

Free trial

Community

Lightweight IDE for Python & Scientific development

DOWNLOAD

Free, open-source

PyCharm (IDE) - - integrated development environment(IDE)

PythonIntroduction - [~/PycharmProjects/PythonIntroduction] - .../lesson1/task1/hello_world.py - PyCharm Educational Edition 1.0

File Edit View Navigate Code Refactor Run VCS Window Help

PythonIntroduction > lesson1 > task1 > hello_world.py

Project

Introduction to Python (PythonIntroduction)
Introduction
Our first program
hello_world.py

hello_world.py

Task Text

Traditionally the first program you write in any programming language is "Hello World!".

Introduce yourself to the World.

Congratulations!

Course Description

Introduction to Python

Author: PyCharm

Introduction course to Python

0 of 10 lessons completed

1 of 50 tasks completed

49 of 50 tasks left

```
1 print("Hello, world! My name is pythonworld")
```

Run Output

```
Hello, world! My name is pythonworld
```

Python Console 4: Run

2:1 n/a UTF-8

4. Applications and examples of the R and Python programming languages

R is used in Google for:

Parallel statistical prediction on big data –

-to improve the effectiveness of Google's online advertising.

- study the effectiveness of search advertising in Google (so, with R, it was found that search advertising gives an additional 89% of web traffic)

Where is Python used?

- Google uses Python in its search engine and pays for the work of the Creator of Python-Guido van Rossum
- Companies such as Intel, Cisco, Hewlett-Packard, Seagate, Qualcomm, and IBM use Python to test hardware
- YouTube's video sharing service is largely implemented in Python
- NSA uses Python to encrypt and analyze intelligence
- JPMorgan Chase, UBS, Getco and Citadel use Python to predict the financial market
- The popular program BitTorrent for file sharing in peer to peer networks is written in Python
- Google's popular App Engine web framework uses Python as an application programming language
- NASA, Los Alamos, JPL, and Fermilab use Python for scientific computing.

Conclusions of the lecture

WE LEARNED:

- What is Big Data
- What does data science do
- Features of the profession Data Scientist
- Software tools for data analysis implementation
- Purpose and benefits of using statistical data processing languages R and Python
- Areas of application of these software tools

Questions for self-control:

1. What features distinguish Big Data from traditional structured data?
2. In what areas of knowledge does Python find its application?
3. What is Rstudio for?