

Delta-4 Intel

Нижний Новгород, 2016



Скриптовые языки программирования

Алексей Бухнин, Intel



Зачем мы здесь?

«Сценарий»

Программа командной оболочки

Связующий слой

Быстрое решение задач

Примеры языков сценариев

Python

Ruby

Lua

JavaScript

PHP

Perl

R

Shell scripts: Bash, Batch

Отличительные особенности

Интерпретируемость

Динамическая типизация

«Сборщик мусора»

Отражение (reflection) и
метапрограммирование

R

- Программная среда вычислений для статистической обработки данных
 - свободная, с открытым исходным кодом
- Язык программирования
 - интерпретируемый
 - аналог S

Особенности

- Реализация большого числа статистических методов
- Доступно более 4000 пакетов расширения
- Эффективные векторизованные функции
- Все объекты хранятся в оперативной памяти
- Принципы ООП основываются на обобщенных функциях
- Не всегда очевидное поведение

Области применения

- Автоматизация анализа данных
 - Автоматизировать все действия
 - Не редактировать результаты генерации
- Воспроизводимые исследования
 - Код на R и пояснения в одном файле
 - Публиковать исходные данные
- Замена коммерческих инструментов

СИНТАКСИС

```
# комментарий
variable <- 'value'
"Can't find it" -> another.variable
is.global.variable <<- TRUE
vector1 <- c(5, 6, 7)
vector1[1]           # [1] 5
vector1[c(2, 3)]    # [1] 6 7
vector1[-3]         # [1] 5 6
vector1[vector1 > 6] # [1] 7
ifelse(vector1 > 5, 1, 0) # [1] 0 1 1
```

СИНТАКСИС

```
PrintAndSum <- function(items=NULL) {  
  if(is.null(items)) {  
    return(0)  
  }  
  for(item in items){  
    print(item)  
  }  
  sum(items)  
}
```

```
result <- PrintAndSum(c(1, 2, as.integer("3")))  
print(result)
```

```
[1] 1  
[1] 2  
[1] 3  
[1] 6
```

Последовательность шагов

- Постановка вопроса исследования
- Загрузка данных
- Обзор данных
- Очистка данных
- Выдвижение гипотезы
- Проверка гипотезы
- Создание отчета

Постановка вопроса исследования

Необходимо определить, возможно ли
различить виды ирисов
(задача об ирисах Фишера)

Ирис щетинистый (*Iris setosa*)



Ирис виргинский (*Iris virginica*)



Ирис разноцветный (*Iris versicolor*)



Обзор данных

```
> names(iris)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length"  
"Petal.Width"  "Species"
```

```
> head(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Обзор данных

```
> str(iris)

'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width  : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width  : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1
1 1 1 1 1 1 ...
```

Обзор данных

```
> table(iris$Species)
```

```
setosa versicolor virginica  
      50         50         50
```

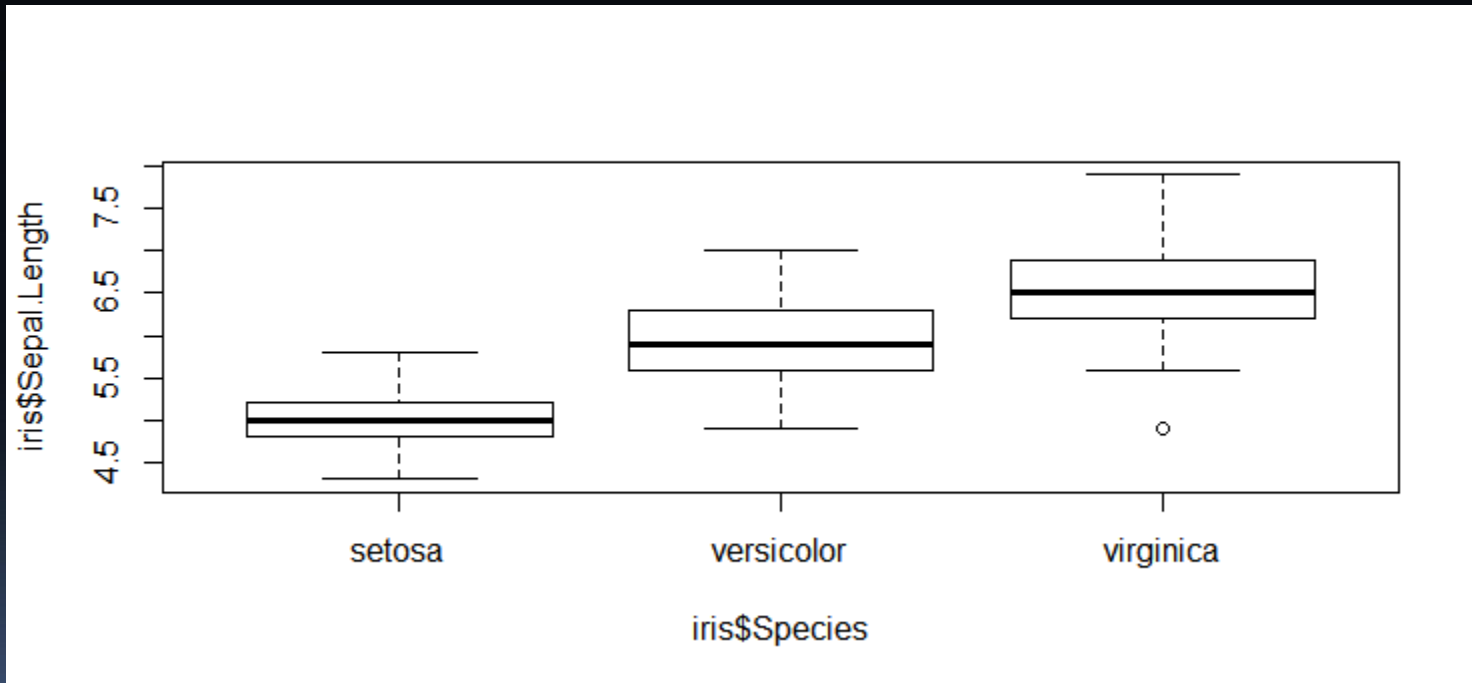

Описательная статистика

```
> summary(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.054	Mean :3.759	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

Графический анализ

```
> plot(iris$Sepal.Length~iris$Species)
```



Выдвижение и проверка гипотезы

Нулевая гипотеза:

Нет статистически значимых различий между группами наблюдений с различными уровнями фактора Species при их сравнении по значениям признаков Sepal.Length, Sepal.Width, Petal.Length, Petal.Width

Автоматизация

Автоматизация анализа

shiny https://abukhnin.shinyapps.io/analyze_groups/

Воспроизводимые исследования (R Markdown)

R Presentation http://rpubs.com/abukhnin/analyze_groups

knitr <http://rpubs.com/abukhnin/stormdata>

Материалы для самообучения

<https://stat.ethz.ch/R-manual/>

<http://manuals.bioinformatics.ucr.edu/home/programming-in-r>

http://www.johndcook.com/blog/r_language_for_programmers/

<http://www.statmethods.net/> (Quick-R)

Шипунов А. и др. "Наглядная статистика. Используем R!"

Patrick Burns "The R Inferno"

<http://google-styleguide.googlecode.com/svn/trunk/Rguide.xml>

http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf (testthat: Get Started with Testing)

Материалы для самообучения

<https://www.coursera.org>

Data Science Specialization

- R Programming
- Reproducible Research
- Developing Data Products

Python

высокоуровневый
интерпретируемый
объектно-ориентированный
многоцелевой язык программирования,
который зачастую используется
для создания сценариев

Положительные стороны Python

Качество программного обеспечения

Высокая скорость разработки

Переносимость программ

Библиотеки поддержки

Недостатки Python

Производительность ниже, чем у компилирующихся языков

Ограничение распараллеливания из-за GIL

Проблемы совместимости версий 2.x и 3.x

Области применения Python

Системное программирование

Интеграция компонентов

Веб-приложения

Приложения баз данных

Быстрое создание прототипов

Программирование научных вычислений

Графический интерфейс

ОСНОВЫ синтаксиса Python

Блоки инструкций выделяются отступами

Конец строки завершает инструкцию

Инструкции на одной строке разделяются ;

После объявления функции, класса,
условного оператора или цикла ставится :

```
for item in items:  
    item += 1; print(item)
```

Рекомендуемая литература

Марк Лутц "Изучаем Python"

"Learning Python"

Марк Лутц "Программирование на Python"

"Programming Python", 2 тома

"Вглубь языка Python"

"Dive Into Python", <http://ru.diveintopython.net/>

Документация к языку

Стив Макконнелл "Совершенный код"

"Code Complete"

Пример интеграции Python и R

Задача: распознавание цифр



5	0	4	1	9	2	1	3		
4	4	6	0	4	5	6	7		
2	0	2	7	1	8	6	4		
1	3	5	9	1	7	6	2		
8	6	3	7	5	8	0	9		
8	7	6	0	9	7	5	7		
2	3	9	4	9	2	1	6		
5	6	7	9	9	3	7	0		

Решение:

R для расчетов

Python Tkinter для GUI

rpy2 для интеграции