北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

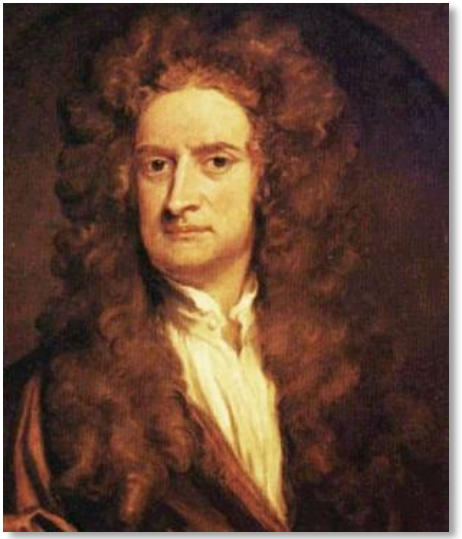
基础编程

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

站在巨人的肩膀上



如果说我看得比别人更远些，那是因为我站在巨人的肩膀上

If I have seen further, it is by standing on the shoulders of giants

站在巨人的肩膀上

人们常说，伟大的科学学科就像是站在其它巨人肩膀上的巨人。人们也说过，软件产业正如站在其他侏儒脚上的侏儒 (Alan Cooper，交互设计之父)

It has been said that the great scientific disciplines are examples of giants standing on the shoulders of other giants. It has also been said that the software industry is an example of midgets standing on the toes of other midgets (Alan Cooper)

站在巨人的肩膀上

The great beauty of R is that you can modify it to do all sorts of things. And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants. (Google, Hal Ronald Varian)

回到Mini小案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

回到Mini小案例

```r
#读取数据
library(readxl)
cjb <- read_excel("data/cjb.xlsx")
View(cjb)
#对数据进行探索性分析
library(tidyverse)
cjb %>%
  select(sx, wlfk) %>%
  ggplot(aes(x = wlfk, y = sx, fill = wlfk)) +
  geom_boxplot()
#数据预处理
as_five_grade_scores <- function(x) {
  cut(x, breaks = c(0, seq(60, 100, by = 10)),
      labels = c("不及格", "及格", "中", "良", "优"))
}
cjb <- cjb %>%
  filter(zcj != 0) %>%#剔除脏数据
  mutate_at(vars(xb, wlfk), factor) %>% #类型转换
  mutate_at(vars(yw:sw), as_five_grade_scores)#数据分箱
View(cjb)
#建模
library(arulesViz)
my_model <- cjb %>%
  select(xb:wlfk) %>%
  apriori(parameter = list(supp = 0.06, conf = 0.8),
  appearance = list(rhs = paste0("wlfk=", c("文科", "理科"))))
#模型评估
inspectDT(my_model)
#可视化
plot(my_model)
```

基础编程

R语言编程人员的定位应该是帅才

需要的是统率三军

充分调动属于某个战队（包）的单兵（函数）

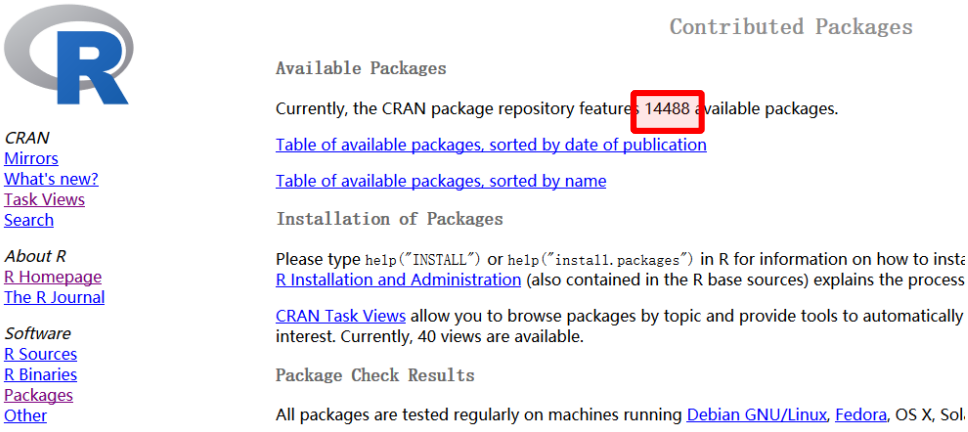
去完成你想要的任务

而不是一切从零开始

****作战单位为单兵*

R编程 = 用别人的包和函数
讲述你自己的故事

大海捞针



如何找到想要的包，野蛮遍历？

大海捞针大法

第1种：看看TV

第2种：逛逛论坛SO

第3种：再不行就SOS

大海捞针大法：TV

第1种：看看TV

官方导航，对R庞大的生态做了一个粗略的划分

第2种：逛逛论坛SO

第3种：再不行就SOS

大海捞针大法：TV

CRAN task views aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic.

Cluster Cluster Analysis & Finite Mixture Models

HighPerformanceComputing High-Performance and Parallel Computing with R

MachineLearning Machine Learning & Statistical Learning

MissingData Missing Data

ModelDeployment Model Deployment with R

NaturalLanguageProcessing Natural Language Processing

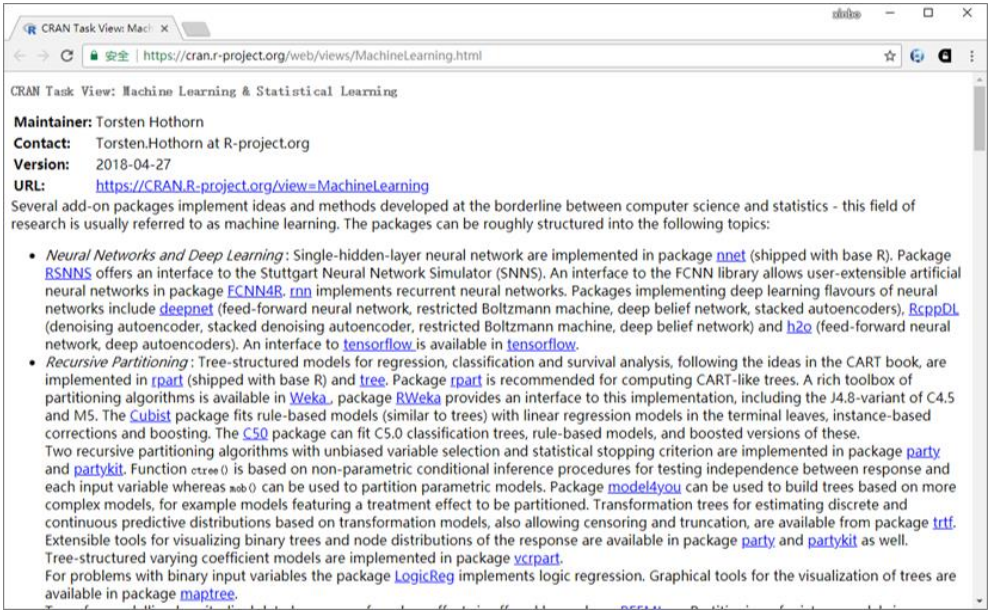
NumericalMathematics Numerical Mathematics

ReproducibleResearch Reproducible Research

Spatial Analysis of Spatial Data

TimeSeries Time Series Analysis

大海捞针大法：TV



https://cran.r-project.org/web/views/MachineLearning.html

大海捞针大法：TV

第1种：看看TV

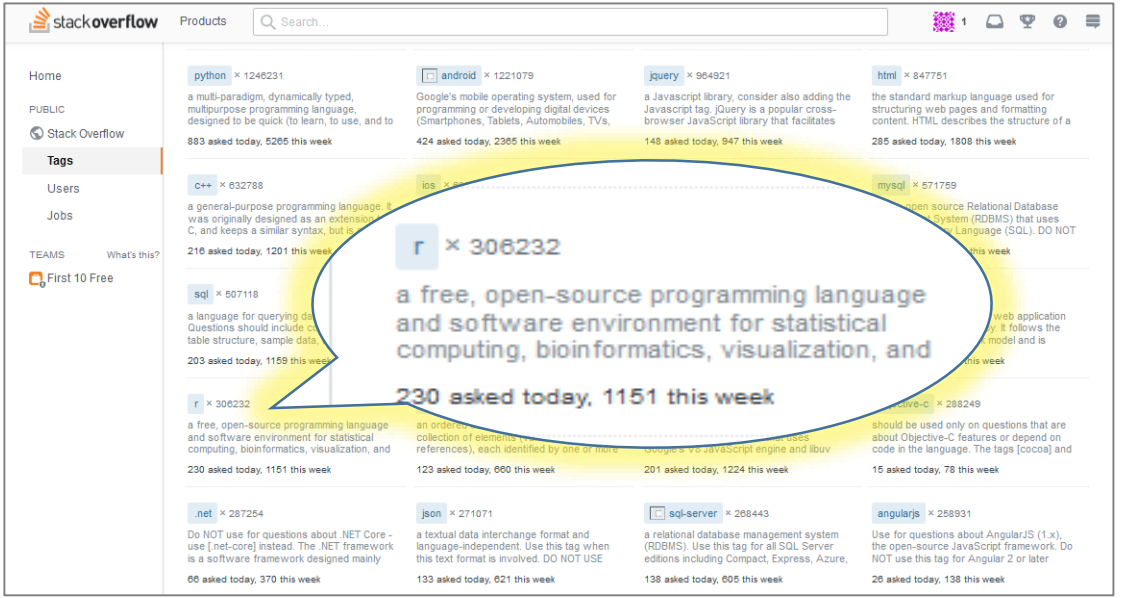
第2种：逛逛论坛SO

http://stackoverflow.com/questions/tagged/r

如果没有它，很多人都不知道怎么编代码

第3种：再不行就SOS

大海捞针大法：SO



https://stackoverflow.com/tags

大海捞针大法：SOS

第1种：看看TV

第2种：逛逛论坛SO

第3种：再不行就SOS

安装sos包

利用findFn()函数进行模糊查找

大海捞针大法：SOS

Help pages for deep+l

file:///C:/Users/byaxb/AppData/Local/Temp/RtmpkR9SGi/file27c817dc69bf.html

findFn Results

call: "x <- sos::findFn(string = 'deep learning')"

For a summary by package, see: "packageSum(x,...)"

See also: vignette('sos')

Id	Count	MaxScore	TotalScore	Package	Function	Date	Score	Description and Link
1	52	368	3678	kerasR	00Index	NA	368	R Interface to the Keras Deep Learning Library
2	52	368	3678	kerasR	ReduceLROnPlateau	NA	241	Reduce learning rate when a metric has stopped improving.
3	52	368	3678	kerasR	Optimizers	NA	80	Optimizers
4	52	368	3678	kerasR	Activation	NA	61	Applies an activation function to an output.
5	52	368	3678	kerasR	ActivityRegularization	NA	61	Layer that applies an update to the cost function based input...
6	52	368	3678	kerasR	AdvancedActivation	NA	61	Advanced activation layers
7	52	368	3678	kerasR	Applications	NA	61	Load pre-trained models
8	52	368	3678	kerasR	AveragePooling	NA	61	Average pooling operation
9	52	368	3678	kerasR	BatchNormalization	NA	61	Batch normalization layer
10	52	368	3678	kerasR	Constraints	NA	61	Apply penalties on layer parameters
11	52	368	3678	kerasR	Conv	NA	61	Convolution layers
12	52	368	3678	kerasR	Cropping	NA	61	Cropping layers for 1D input (e.g. temporal sequence).
13	52	368	3678	kerasR	CSVLogger	NA	61	Callback that streams epoch results to a csv file.
14	52	368	3678	kerasR	Datasets	NA	61	Load datasets
15	52	368	3678	kerasR	decode_predictions	NA	61	Decode predictions from pre-defined imagenet networks

扩展包：机器学习与数据挖掘

	Package	CRAN	Stack Overflow	GitHub
1	caret	1	3	3
2	randomForest	3	4	
3	e1071	2	7	
4	rpart	7	5	
5	nnet	4	9	
6	glmnet	6	8	
7	tree	13	2	
8	party	10	6	
9	arules	9	10	8
10	kernlab	5	17	
11	ROCR	8	16	
12	gbm	14	11	
13	RWeka	12	15	
14	rattle	11	18	
15	earth	18	13	
16	h2o	20	14	1

https://github.com/thedataincubator/data-science-blogs/blob/master/top-r-packages.md (Accessed on May 20, 2018)

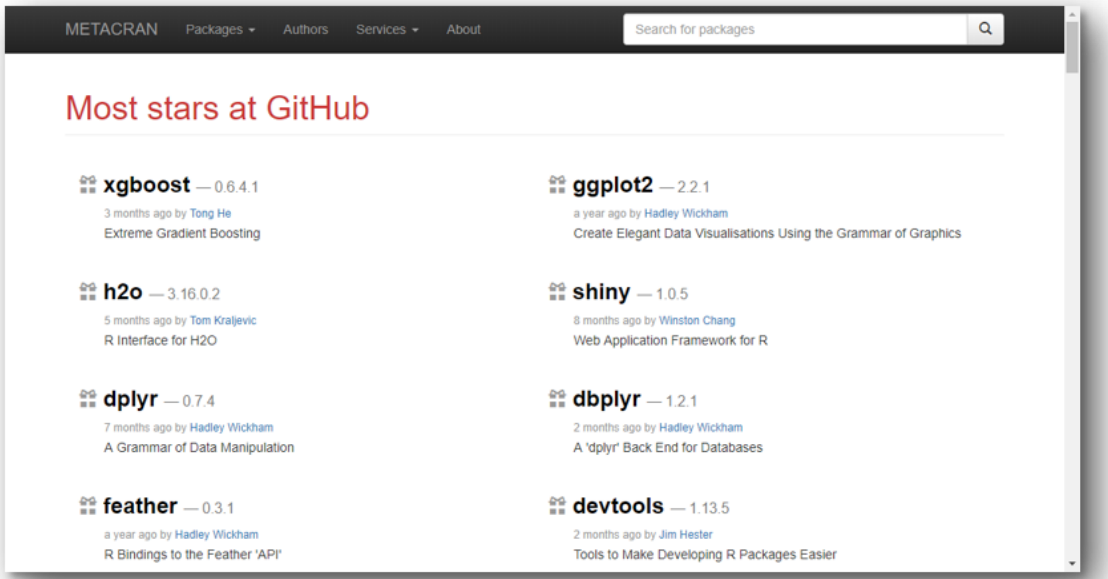
扩展包：机器学习与数据挖掘

```
#caret包所支持的模型library(caret)
library(tidyverse)
modelLookup() %>%
  select(-(2:3)) %>%
  distinct() %>%
  View()
```

	model	forReg	forClass	probModel
1	ada	FALSE	TRUE	TRUE
2	AdaBag	FALSE	TRUE	TRUE
3	adaboost	FALSE	TRUE	TRUE
4	AdaBoost.M1	FALSE	TRUE	TRUE
5	amdai	FALSE	TRUE	TRUE
6	ANFIS	TRUE	FALSE	FALSE
7	avNNet	TRUE	TRUE	TRUE
8	awnb	FALSE	TRUE	TRUE
9	awtan	FALSE	TRUE	TRUE
10	bag	TRUE	TRUE	TRUE
11	bagEarth	TRUE	TRUE	TRUE
12	bagEarthGCV	TRUE	TRUE	TRUE
13	bagFDA	FALSE	TRUE	TRUE
14	bagFDAGCV	FALSE	TRUE	TRUE
15	bam	TRUE	TRUE	TRUE

Showing 1 to 15 of 237 entries

扩展包：Most stars at GitHub



https://www.r-pkg.org/starred

扩展包地图：机器学习与数据挖掘

=====数据读取====

readr：实现表格数据的快速导入

readxl：读取Microsoft Excel电子表格数据

rvest：网页数据抓取包

RMySQL：用于连接MySQL数据库的R包

=====探索性分析=====

Hmisc：统计描述

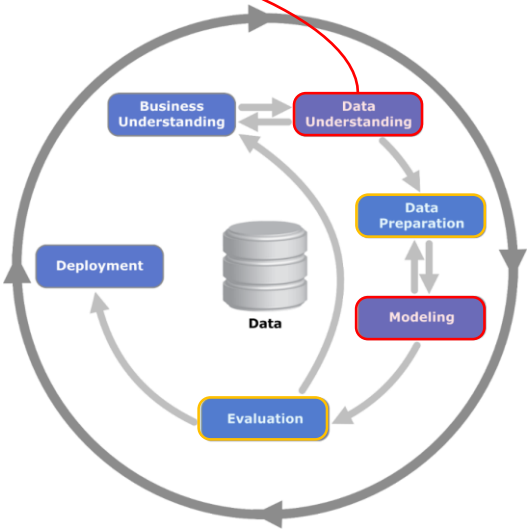
ggplot2及其扩展

GGally：绘制散点图矩阵

leaflet：绘制交互式地图

plotly：交互式绘图包

可视化：Task Views > Graphics



扩展包地图：机器学习与数据挖掘

====数据处理、类型转换等====

tidyverse: 长款变换、更改、分组统计等

data.table：用于快速处理大数据集

stringr、 stringi：一个字符串处理工具集

lubridate：用于处理日期时间数据

xts：时间序列

Matrix：著名的稀疏矩阵包

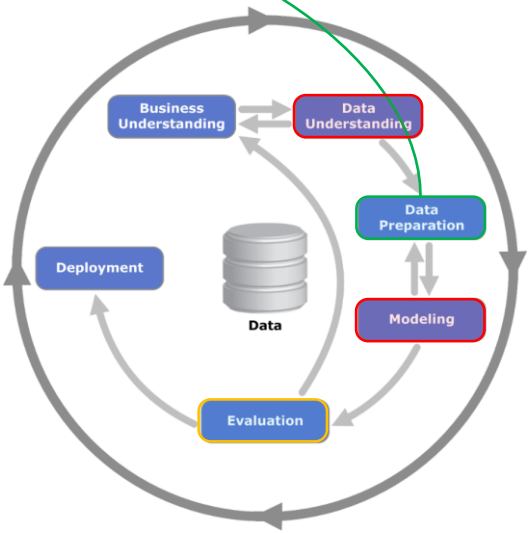
====缺失值处理==== Task Views > MissingData

VIM：缺失值处理

mice：缺失值处理

====降维====

psych、vegan、MASS：主成分分析、因子分析、MDS



扩展包地图：机器学习与数据挖掘

=====典型算法模型====

caret：分类与回归框架包

mlr：框架包

randomForest：随机森林

rpart：树模型

arules：频繁项集与关联

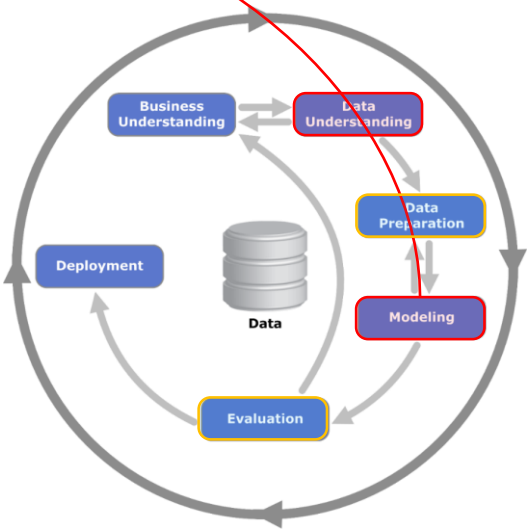
e1071：比较流行的机器学习包

cluster：聚类分析

factoextra：聚类分析及其可视化

神经网络与深度学习：nnet、RSNNS、deepnet、RcppDL、h2o、tensorflow等

更多内容参见：Task Views > Machine Learning



扩展包地图：机器学习与数据挖掘

====模型选择与验证====

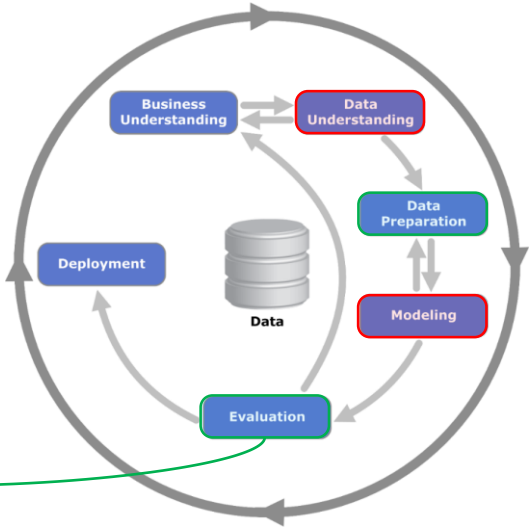
caret：自带多种评估指标，支持格子点搜索

e1071：自带多种评估指标

ROCR：通过绘图来可视化分类器的综合性能

MLmetrics：提供了多种指标

NBclust：聚类分析性能评估



扩展包地图：机器学习与数据挖掘

=====结果交流====

rmarkdown ： 可重复性报告和动态文档

knitr：用于在PDF和HTML文档中嵌入R代码块

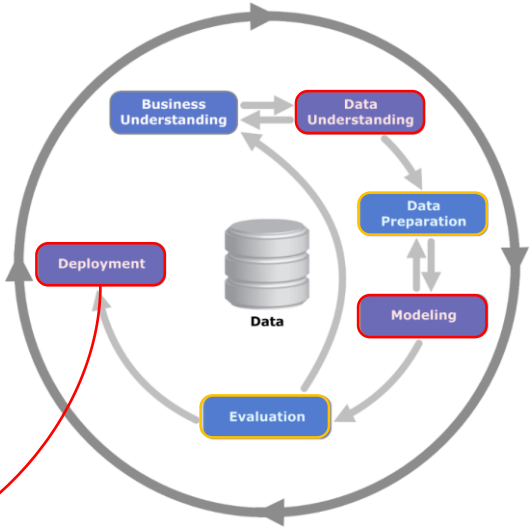
flexdashboard：创建仪表盘

=====模型部署====

pmml：转换成PMML

plumber：web API接口

shiny：交互式web应用



利用好帮助文档

```
xm <- c("周黎", "汤海明", "舒江辉", "翁柯", "祁强", "湛容")
?c
```

c {base} R Documentation

Combine Values into a Vector or List

Description

This is a generic function which combines its arguments.

The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.

Usage

```
## S3 Generic function
c(...)

## Default S3 method:
c(..., recursive = FALSE, use.names = TRUE)
```

Arguments

...	objects to be concatenated.
recursive	logical. If `recursive = TRUE`, the function recursively descends through lists (and pairlists) combining all their elements into a vector.
use.names	logical indicating if `names` should be preserved.

利用好帮助文档

```
xm <- c("周黎", "汤海明", "舒江辉", "翁柯", "祁强", "湛容")
?c
```

Details

The output type is determined from the highest type of the components in the hierarchy NULL < raw < logical < integer < double < complex < character < list < expression. Pairlists are treated as lists, whereas non-vector components (such names and calls) are treated as one-element lists which cannot be unlisted even if `recursive = TRUE`.

Note that factors are treated only via their internal integer codes; one proposal has been to use

```
c.factor <- function(..., recursive=TRUE) unlist(list(...), recursive=recursive)
```

if factor concatenation by `c()` should give a factor.

`c` is sometimes used for its side effect of removing attributes except names, for example to turn an array into a vector. `as.vector` is a more intuitive way to do this, but also drops names. Note that methods other than the default are not required to do this (and they will almost certainly preserve a class attribute).

This is a primitive function.

Value

`NULL` or an expression or a vector of an appropriate mode. (With no arguments the value is `NULL`.)

S4 methods

This function is S4 generic, but with argument list `(x, ...)`.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

See Also

利用好帮助文档

```
xm <- c("周黎", "汤海明", "舒江辉", "翁柯", "祁强", "湛容")
?c
```

See Also

unlist and as.vector to produce attribute-free vectors.

Examples

```
c(1,7:9)
c(1:5, 10.5, "next")

## uses with a single argument to drop attributes
x <- 1:4
names(x) <- letters[1:4]
x
c(x)          # has names
as.vector(x)  # no names
dim(x) <- c(2,2)
x
c(x)
as.vector(x)

## append to a list:
ll <- list(A = 1, c = "C")
## do *not* use
c(ll, d = 1:3) # which is == c(ll, as.list(c(d = 1:3)))
## but rather
c(ll, d = list(1:3))  # c() combining two lists

c(list(A = c(B = 1)), recursive = TRUE)

c(options(), recursive = TRUE)
c(list(A = c(B = 1, C = 2), B = c(E = 7)), recursive = TRUE)
```

[Package *base* version 3.5.0 Index]

谢谢聆听

Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics