天善智能
TIANSHAN SOFT
专注于商业智能BI和大数据的垂直社区平台

R语言基础知识

谢佳标（Daniel.xie）

数据分析/挖掘前景

2015年1月，Linkedin对全球超过3.3亿用户的工作经历和技能进行分析，公布2014年最受雇主喜欢、最炙手可热的25项技能，统计分析和数据挖掘位列榜首。

热

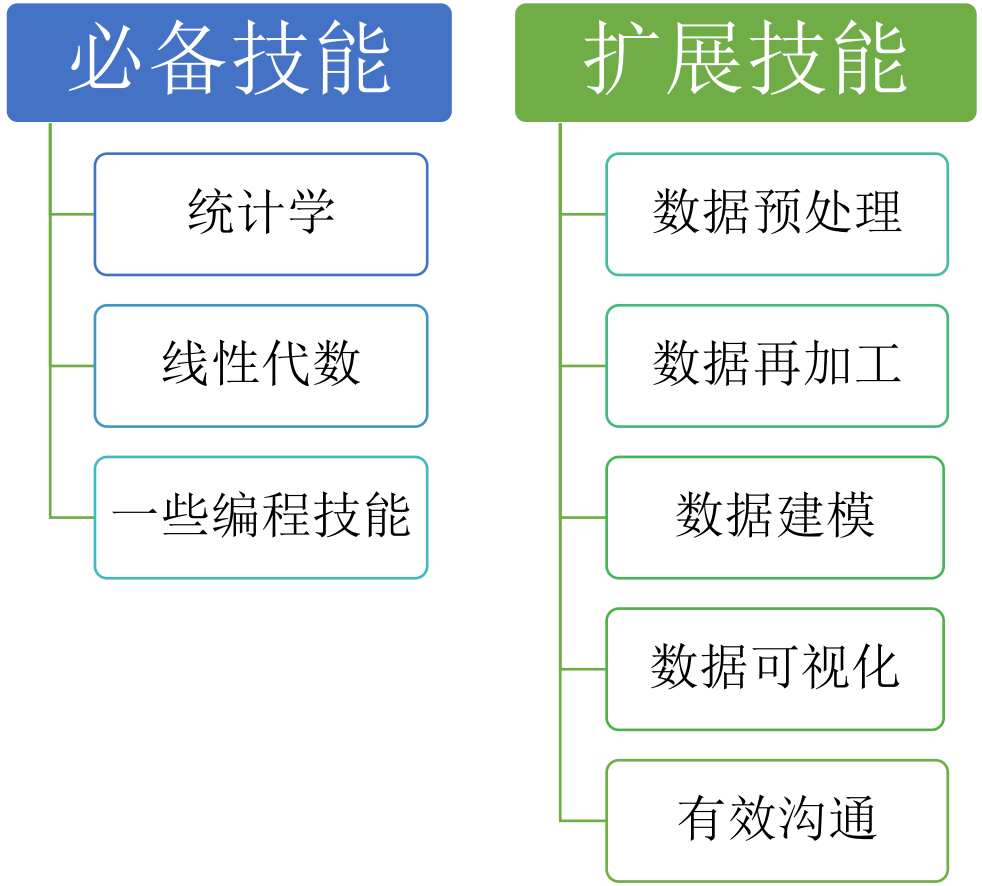
据艾瑞的研究报告，未来与数据分析相关的就业岗位会在1000万左右，而目前来说国内的合格的数据分析师不足5万左右

缺

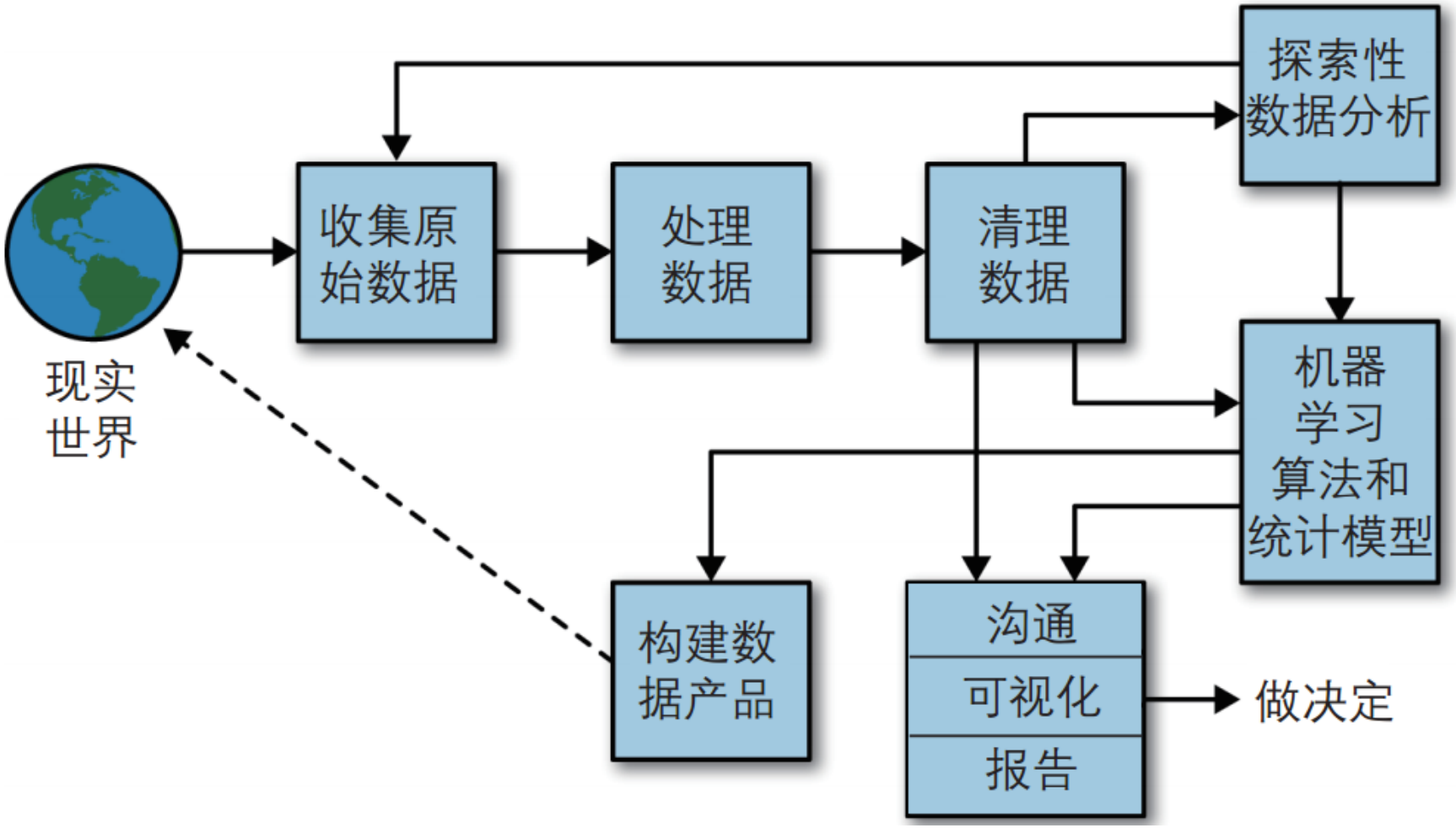
企业也希望能在找到一个合格的数据分析，希望在互联网与大数据时代，把握整个企业在市场上的走向

专

如何成为一名的数据从业者

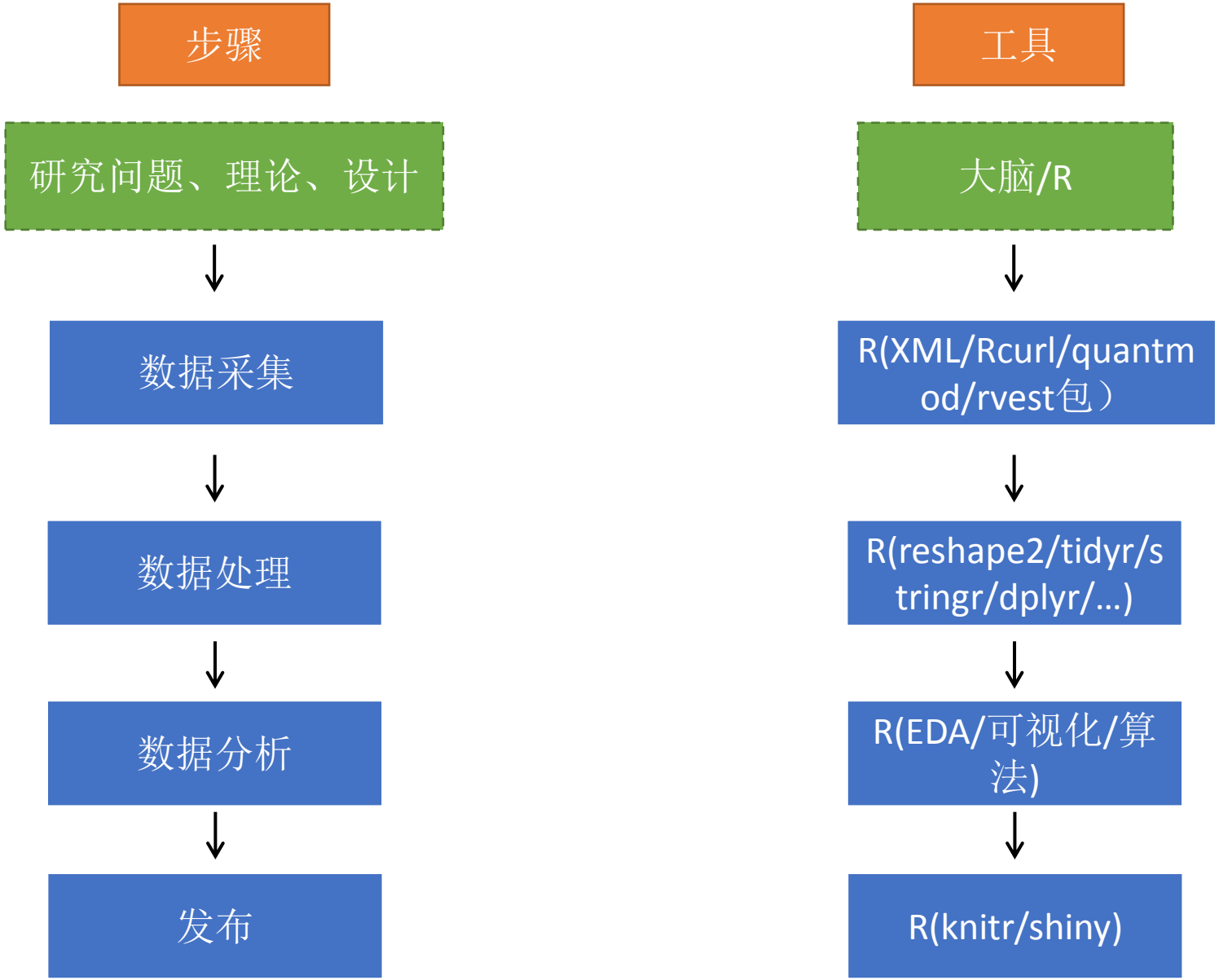


数据科学的工作流程



摘自《数据科学实战》

使用R进行数据挖掘



R快速入门

软件安装

- **Windows下安装R、Rstudio**

方法：从http://www.r-project.org/网站上下载R 安装文件

从http://www.rstudio.com/网站上下载RStudio安装文件

- **linux下安装R、Rsutdio**

方法：执行sudo apt-get install r-base-dev 安装R

执行wget https://download1.rstudio.org/rstudio-1.0.136-amd64.deb

sudo gdebi rstudio -1.0.136-amd64.deb安装rstudio

基本操作

- 查找帮助
- 工作空间
- 包的使用
- 数据读入

安装R包

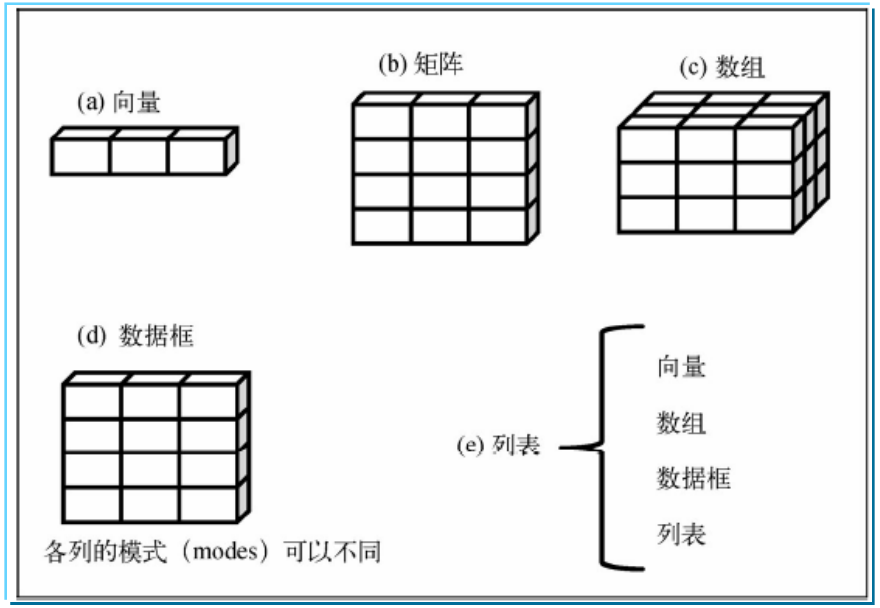
- install.packages()
- devtools::install_github()
- RCMD INSTALL "xxx.tar.gz"
- 本地安装(通过窗口操作)

数据对象

- 向量(vector)
- 列表(list)
- 矩阵(matrix)
- 数据框(data.frame)

数据对象

- R拥有许多用于存储数据的对象类型，包括向量、矩阵、数组、数据框和列表。它们在存储数据的类型、创建方式、结构复杂度，以及用于定位和访问其中个别元素的标记等方面均有所不同。
- 在R中，对象(object)是指可以赋值给变量的任何事物，包括常量、数据结构、函数甚至图形。
- 数据框(data frame)是R中用于存储数据的一种结构：列表示变量，行表示观测。在同一个数据框中可以存储不同类型(如数值型、字符型)的变量。数据框将是你用来存储数据集的主要数据结构。
- 因子(factor)是名义型变量或有序型变量。它们在R中被特殊地存储和处理。



数据的创建

- 通俗地说，对象类型是指R语言组织和管理内部元素的不同方式。数据类型则描述了一个变量内元素取值的类型。例如，逻辑类型数据的取值是TRUE和FALSE，而数值类型的取值是实数。不同对象类型元素取值的数据类型如下表所示：

对象类型	数据类型	是否允许出现不同数据类型
向量	数值型、复数型、字符型、逻辑型	不允许
因子	数值型、复数型、字符型、逻辑型	不允许
数组	数值型、复数型、字符型、逻辑型	不允许
矩阵	数值型、复数型、字符型、逻辑型	不允许
数据框	数值型、复数型、字符型、逻辑型	相同列内元素，其数据类型必须相同； 不同列之间的数据类型可以不同
列表	数值型、复数型、字符型、逻辑型	任何元素的数据类型均可不同
时间序列	数值型、复数型、字符型、逻辑型	不允许

- 对于未知类型的对象，在R中有3个函数可以查看对象的类型：class()、mode()、typeof()。

向量

- 向量是以一维数组的方法管理数据的一种对象类型。可以说向量是R语言中最基本的数据类型，很多算法函数都是以向量的形式输入的。
- 向量可以是字符型、逻辑值型(T、F)、数值型和复数型。
- 在大多数情况下，使用长度大于1的向量。可以在R中使用c( )函数和相应的参数来创建一个向量。
- 一个对象的长度是它含有元素的数量，可以用length( )函数来获取
- 一个向量的所有元素都必须属于相同的模式。如果不是，R将强制执行类型转换。
- R语言最强大的方面之一就是函数的向量化。这些函数可以直接对向量的每个元素进行操作。

矩阵和数组

■向量vector用于描述一维数据，是R语言中最基础的数据结构形式，然而在很多情况下，数据是以二维甚至多维的形式存在的。

■利用矩阵matrix可以描述二维数据，和向量相似，其内部元素可以是实数、复数、字符、逻辑型数据。矩阵matrix使用两个下标来访问元素，A[i,j]表示矩阵A第i行、第j列的元素。

■多维数组array可以描述多维数据。array有一个特征属性叫维数向量（dim属性），它的长度是多维数组的维数，dim内的元素则是对应维度的长度。

■矩阵是数组的特殊情况，它具有两个维度。

列表和数据框

■列表list和数据框data.frame也是一个二维数据，其中向量vector、多维数组array以及矩阵matrix存储的元素，其数据类型是唯一的。列表和数据框内每列元素的数据类型可以不同，列表内的长度也可以不同。

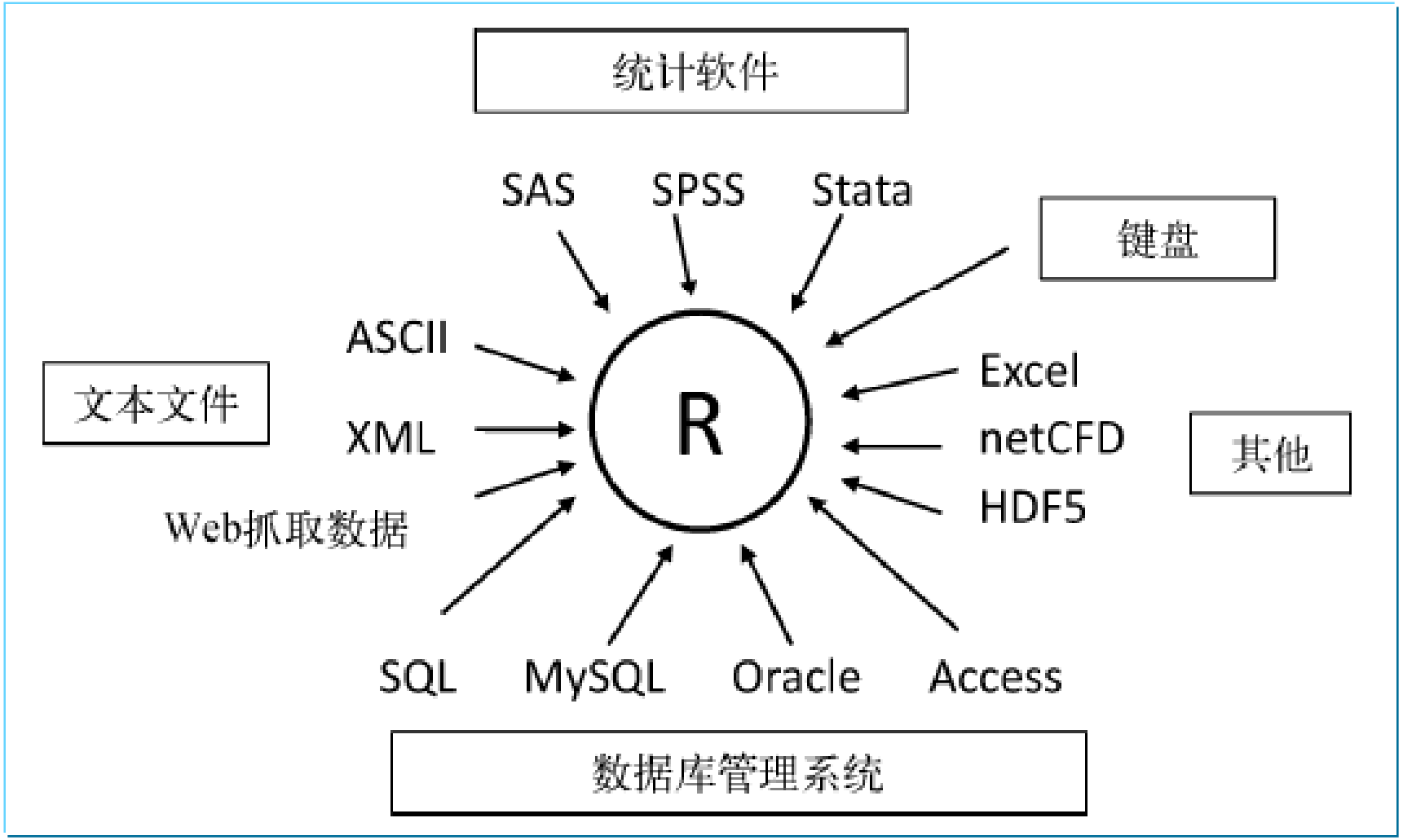
■一般地，在使用R语言进行数据分析和挖掘的过程中，向量和数据框的使用频率是最高的，list则在存储较复杂的数据时作为数据对象类型。

■list（ ）可以用于创建列表对象。

■data.frame（ ）函数可以直接把多个向量建立为一个数据框，并为列设置名称。

可供R导入的数据源

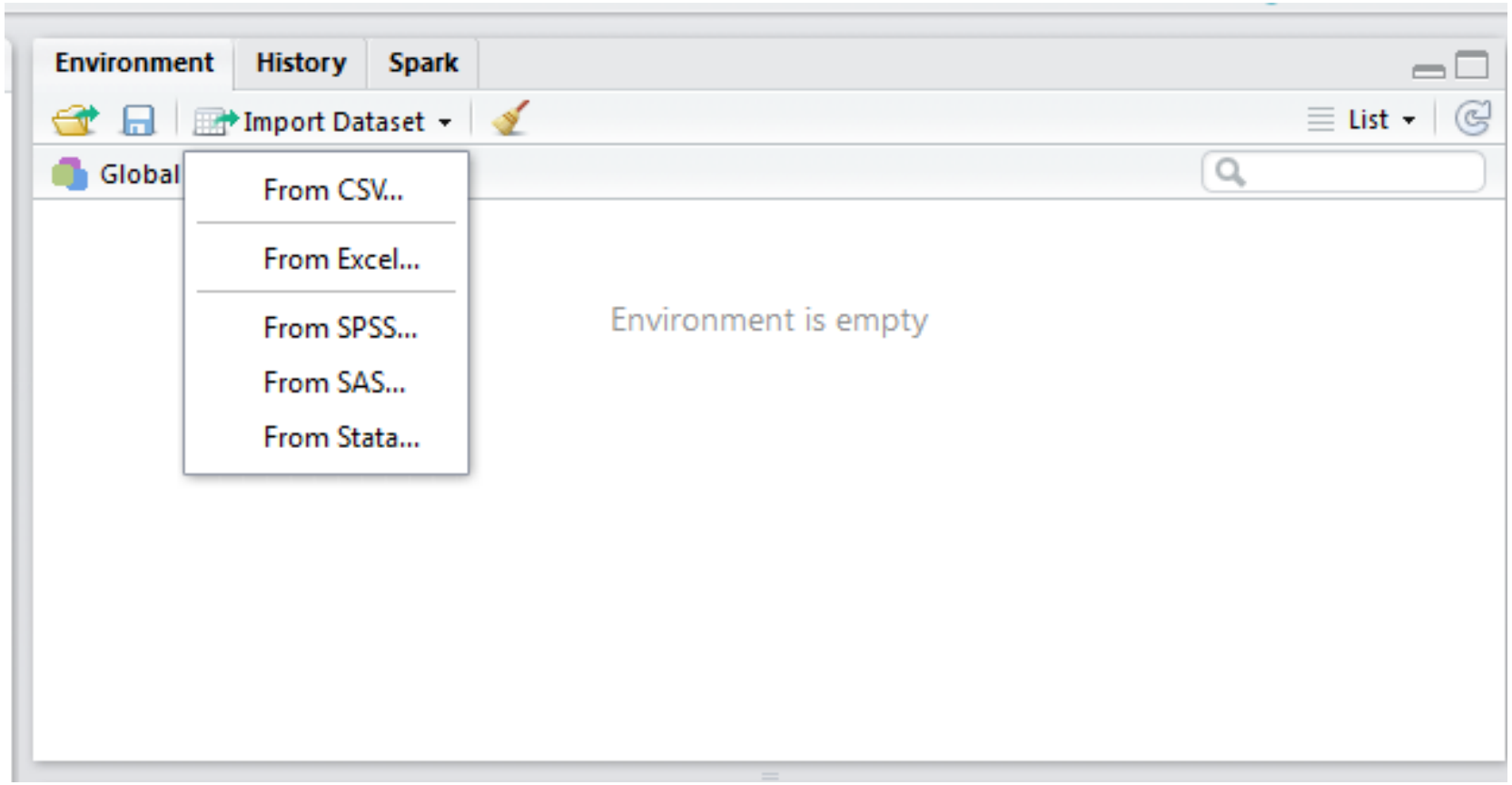
- R可以从键盘、文本文件、Microsoft Excel和Access、流行的统计软件、特殊格式的文件，以及多种关系型数据库中导入的数据。



利用Rstuido进行数据的导入

■在了解R的数据结构以后，接下来要做的就是导入数据。R暂时没有很好用的可视化的数据导入工具，所有需要使用命令来导入导出数据。

■如果使用Rstudio编辑器，可以使用其提供的简单的数据导入功能：



常用的读取指令read

■R最常用的读取文本文件(ASCII）的指令是read.table( )，它是读取矩阵格子状数据最为便利的方式。

read.table()指令的格式如下：

```
read.table(file, header = FALSE, sep = "", quote = "\"'", dec = ".", row.names, col.names,
    as.is = !stringsAsFactors, na.strings = "NA", colClasses = NA, nrows = -1,
    skip = 0, check.names = TRUE, fill = !blank.lines.skip, strip.white = FALSE,
    blank.lines.skip = TRUE, comment.char = "#")
```

参数名	含　义
file	要读取的数据文件名称
header	逻辑值，TRUE 表示文件的第一行包含变量名，默认为 FALSE
sep	文件中字段的分隔符，默认为 sep=" "，表示分隔符是空格
quote	设置如何引用字符型变量。默认情况下，字符串可以被引号"或'括起，如果没有设定分隔字符，引号前面加\，即 quote="\"
dec	设置用来表示小数点的字符
row.names	向量的行名，默认为 1,2,3,…
col.names	向量的列名，默认为 V1,V2,V3,…
na.strings	赋给缺失数据的值（NA）
skip	开始读取数据前跳过的数据文件的行数
strip.white	是否消除空白字符
blank.lines.skip	是否跳过空白行

导入Excel数据

- 读取一个Excel文件的最好方式，就是在Excel中将其导出为一个都好分割文件(csv)，并使用read.csv( )的方式将其导入R中。
- 在Windows系统中，可以使用RODBC包来访问Excel文件，或直接用xlsx包、XLConnect包和readxl包来访问Excel2007文件。

```
> # 利用xlsx包读取EXcel数据
> library(xlsx)
> file<-'sample.xlsx'
> res <- read.xlsx(file,1)
> res
  FirstName LastName Income
1       Joe    Smith  1e+05
2      Mike    Steel  2e+03
3       Liv    Storm  8e+03
> detach(package:xlsx)
> # 利用XLConnect包读取Excel数据
> library(XLConnect)
> wb <- loadWorkbook("sample.xlsx")
> xldf<-readWorksheet(wb,sheet=getSheets(wb)[1])
> xldf
  FirstName LastName Income
1       Joe    Smith  1e+05
2      Mike    Steel  2e+03
3       Liv    Storm  8e+03
>
```

其他文件读取

- 由于某些原因，可能需要从其他格式的文件中读入数据，比如SAS的数据文件、SPSS的数据文件等。下表列出了foreign包中读取外部数据的函数。

函数	描述
read.arff	从ARFF文件中读取文件，著名的数据挖掘开源软件weka的数据就是这种格式
read.dbf	读取DBF文件，DBF文件就是数据库文件
read.dta	读取Stata中的数据集
read.epiinfo	读取Epi Info的数据集
read.mtp	读取Minitab中的数据集
read.octave	读取Octave的文本数据
read.spss	读取SPSS的数据文件
read.ssd	读取SAS的永久数据集
read.systat	读取Systat格式的数据

访问数据库管理系统

■R中有多种面向关系型数据库管理系统(DBMS)的接口，包括SQL Server、Access、MySQL、Oracle、DB2等。其中一些包通过原生的数据库驱动来提供访问功能，另一些则是通过ODBC或JDBC来实现访问的。使用R来访问存储在外部数据库中的数据是一种分析大数据集的有效手段，并且能够发挥SQL和R。

1. ODBC接口

■在R中通过RODBC包访问一个数据库也许是最流行的方式，这种方式允许R连接到任意一种拥有ODBC驱动的数据库，这包含了上面所列的所有数据库

2. DBI相关包

■DBI包为访问数据库提供了一个通用且一致的客户端接口。构建于这个框架之上的RJDBC包提供了通过JDBC驱动访问数据库的方案。使用时请确保安装了针对你的系统和数据库的必要JDBC驱动。其他有用的、基于DBI的包有RMySQL、 ROracle、 RPostgreSQL和RSQLite。

案例演示

1. 案例一：RODBC在windows上的安装及演示

2. 案例二：RMySQL在windows上的安装及演示

读取网络数据

- 网络上的数据，可以通过所谓Web数据抓取（Webscraping）的过程，或对应用程序接口（application programming interface， API）的使用来获得。

- 一般地说，在Web数据抓取过程中，用户从互联网上提取嵌入在网页中的信息，并将其保存为R中的数据结构以做进一步的分析。比如说，一个网页上的文字可以使用函数readLines()来下载到一个R的字符向量中，然后使用如grep()和gsub()一类的函数处理它。对于结构复杂的网页，可以使用rvest包、RCurl包和XML包来提取其中想要的信息。

案例演示

1. 案例一：通过XML包和rvest包爬取网上表格
2. 案例二：通过readLines函数会rvest包爬取团购网信息