R 在大规模数据整理及自动化报告方面的应用

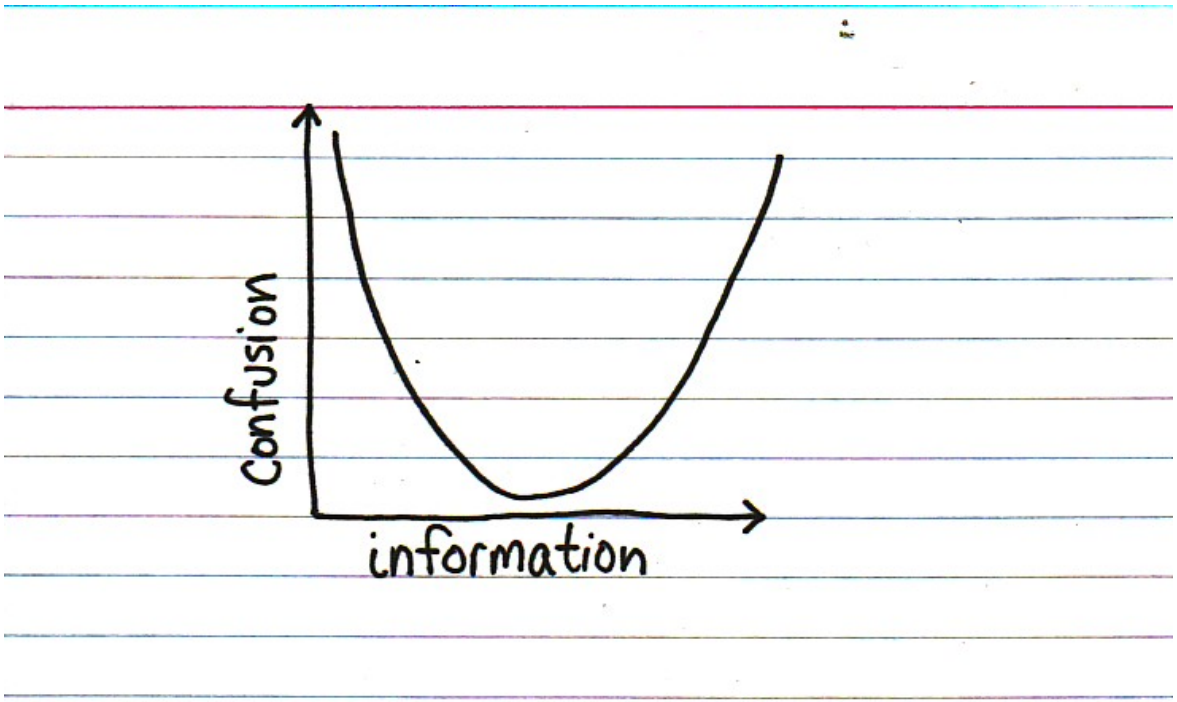
刘思喆

China Lottery Online Co.,Ltd 2009

2009.12.05

提纲

为什么需要数据整理？



Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented.

When dealing with large data sets in R

1. 内存限制
2. 算法问题
3. 索引问题

Memory limitations

1. 32 位操作系统上，可利用的最大内存为 4GB
2. 64 位操作系统的内存也是 “有限” 的，并且费用 · · ·
3. Moore’s Law 似乎已经达到上限

大规模数据整理的商业解决方案

ETL（Extraction，Transformation，Loading）

- 专业的 ETL 工具
- SQL 编程
- ETL 工具和 SQL 相结合

如果用 R 呢

RODBC is a mature and much-used platform for interfacing R to database systems.

sqldf is an R package for performing SQL statements on R data frames, optimized for convenience.

RODBC

sqlQuery Submit an SQL query to an ODBC database, and retrieve the results.

sqlSave Write or update a table in an ODBC database.

RODBC

sqlQuery Submit an SQL query to an ODBC database, and retrieve the results.
The term 'query' includes any valid SQL statement including table creation, alteration, updates etc as well as 'SELECT's.

sqlSave Write or update a table in an ODBC database.

sqldf

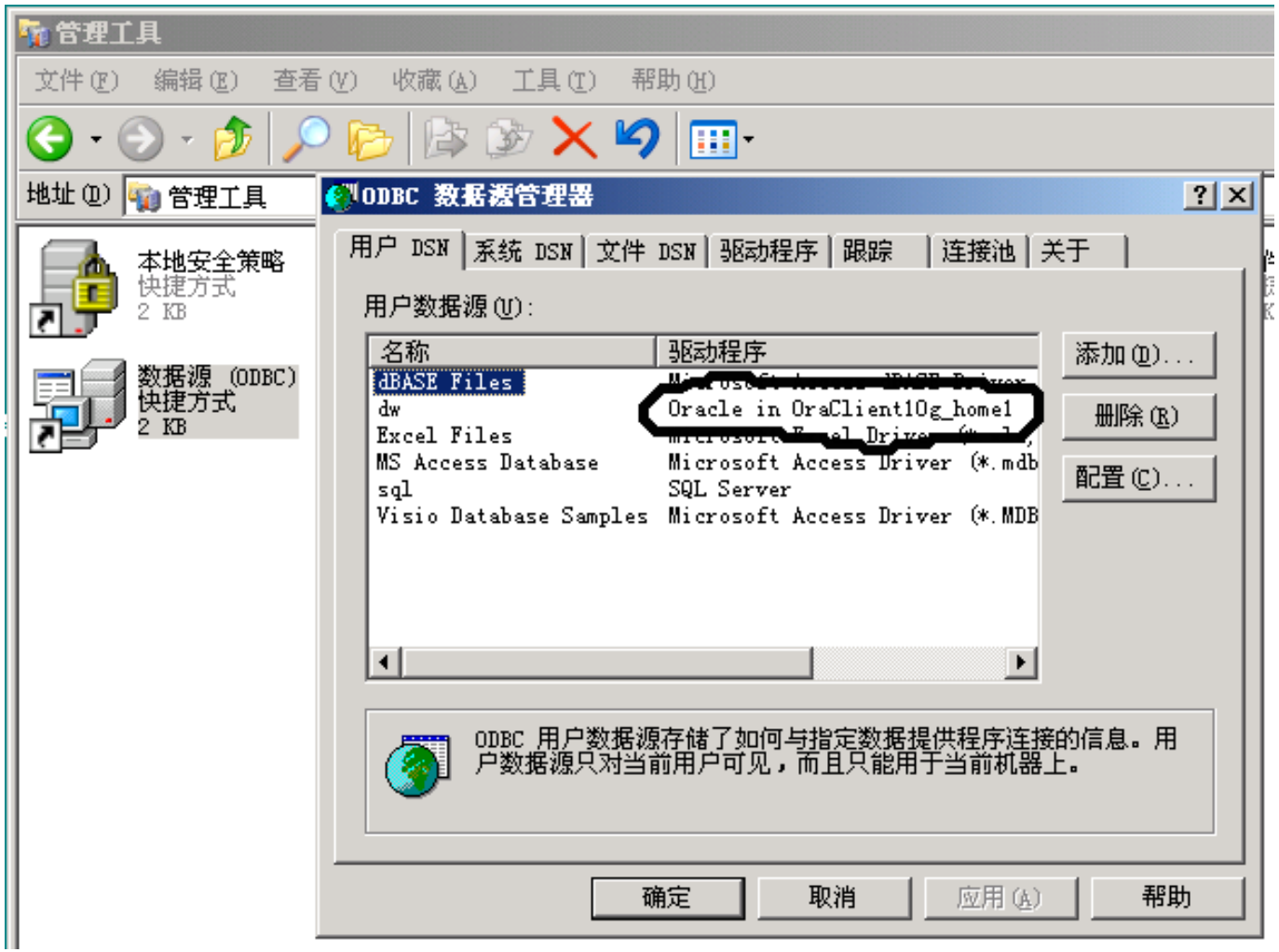
- as an alternate syntax for data frame manipulation
- learning SQL if you know R;learning R if you know SQL
- reading portions of large files into R without reading the entire file
- reading random selection of rows from a file

ODBC Driver Manager and Driver

OS	ODBC Driver Manager
Mac OS X	iodbc
Linux	iodbc,unixODBC
Windows	Microsoft Data Access Components

The driver depends on your data base.

Installation of RODBC



```
setwd("D:/work/oracle")
channel <- odbcConnect("dw",uid = "Ruser",pwd = "Ruser")
script  <- readLines('script.sql')
result  <- sqlQuery(channel,paste(script,collapse = ''))
```

提纲

数据环境

- 中福在线共有 7 款游戏
- 2009 年 8 月份数据大约 3 亿条销售数据
- 平均每天的数据量达到 1000 万条记录
- 大约 20 个业务字段

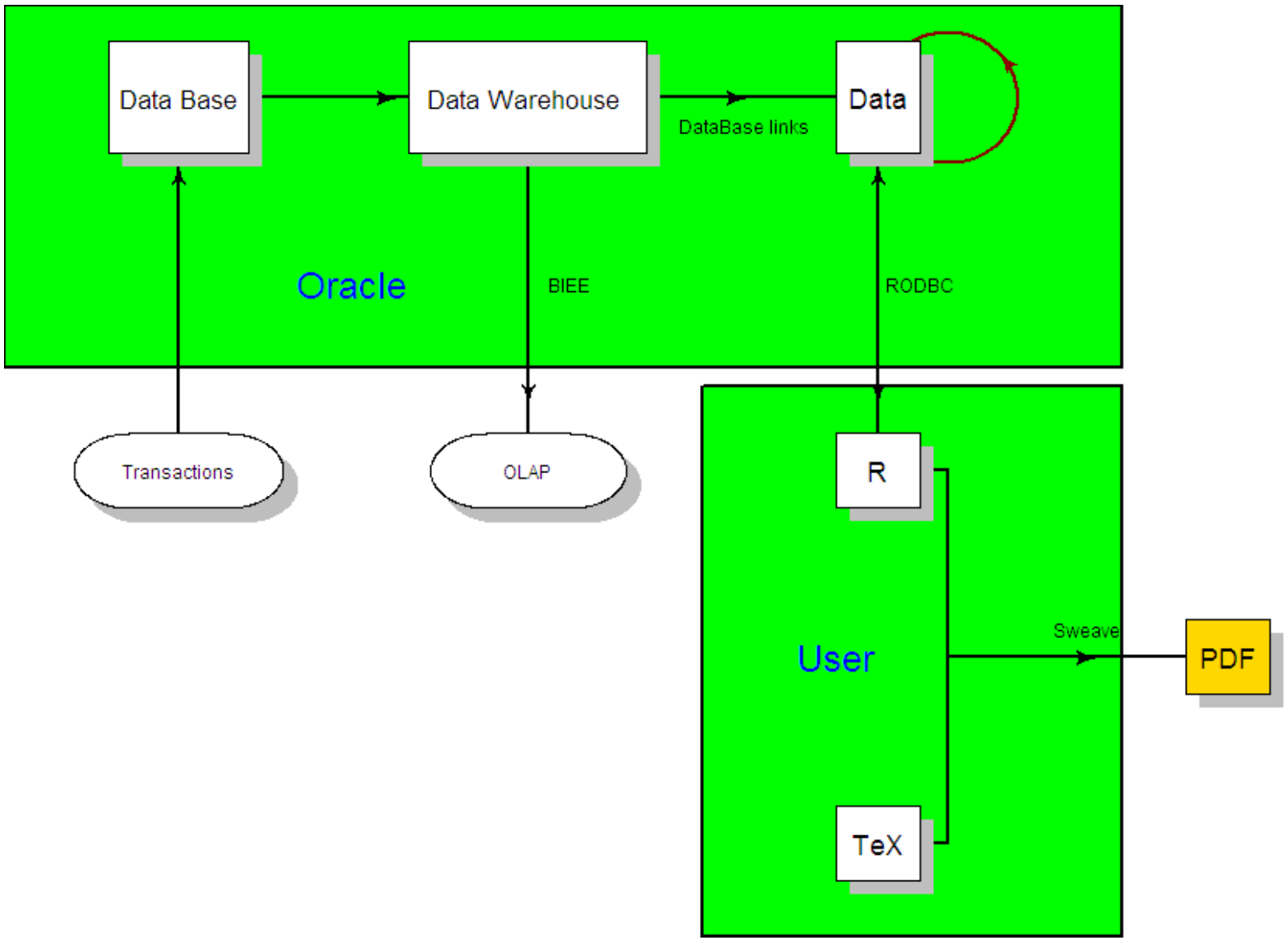
数据环境

- 中福在线共有 7 款游戏
- 2009 年 8 月份数据大约 3 亿条销售数据
- 平均每天的数据量达到 1000 万条记录
- 大约 20 个业务字段

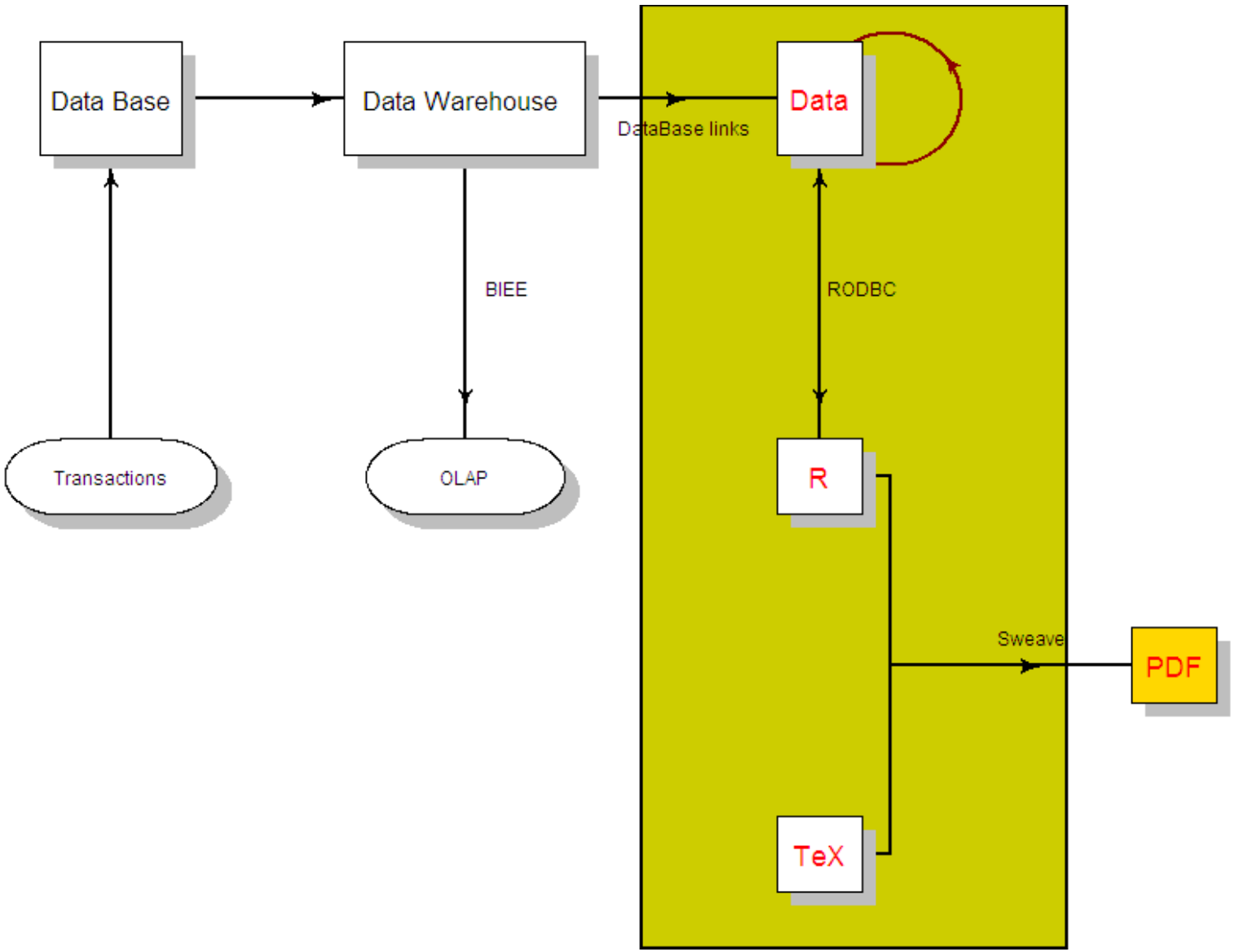
任务

描绘在线人数走势图

系统构架

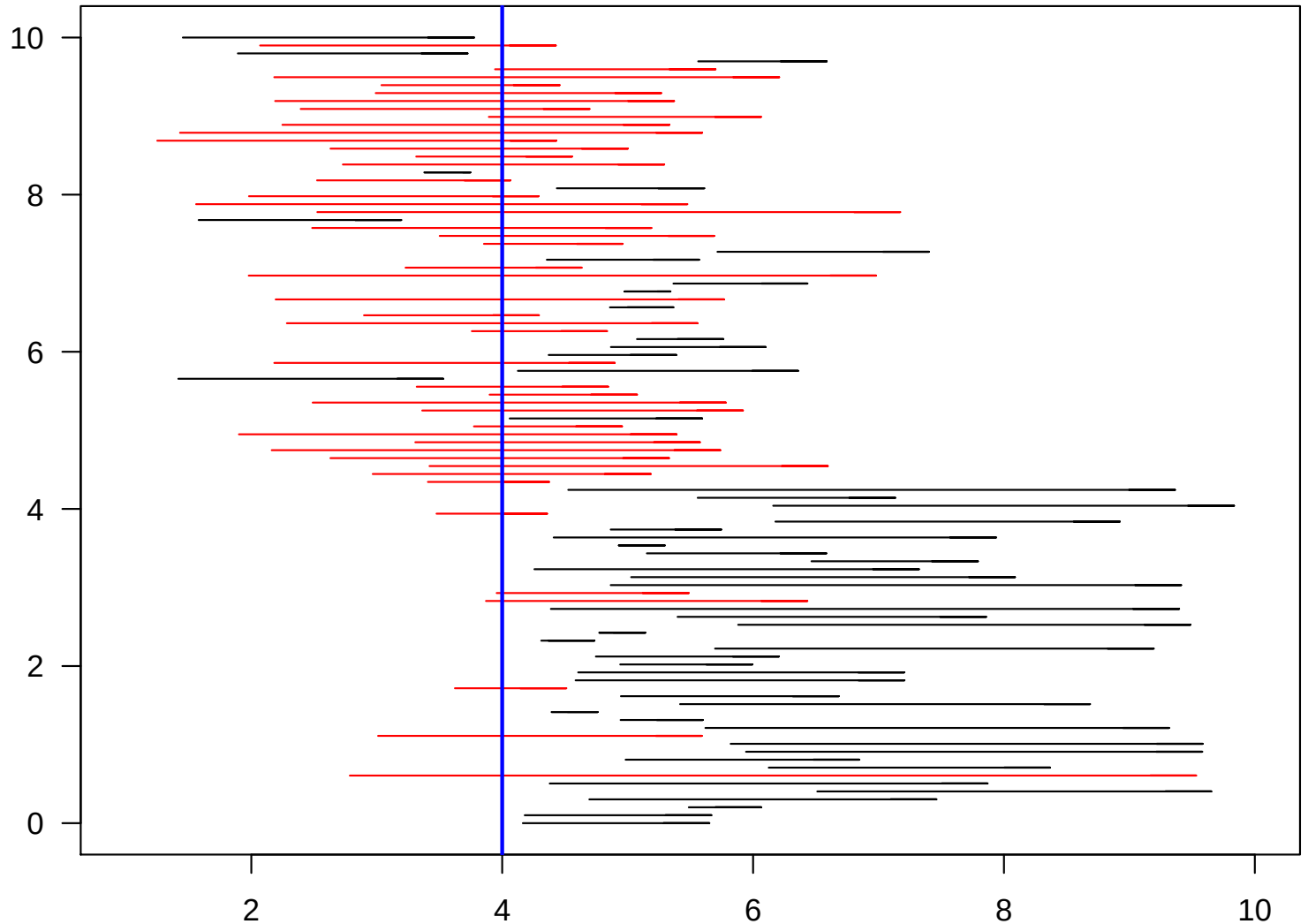

Data Base
Data Warehouse
DataBase links
Data
Oracle
BIEE
RODBC
Transactions
OLAP
R
User
Sweave
PDF
TeX

系统构架


Data Base
Data Warehouse
DataBase links
Data
BIEE
RODBC
Transactions
OLAP
R
Sweave
PDF
TeX

模型构建

模型构建



结果演示

300000000 条数据 =?

结果演示

300000000 条数据 =?

Wait· · ·

提纲

R 灵活的扩展性使得我们可以将任意需要的数据载入至 R ，并通过 Sweave 技术将其生成报告。

Sweave 提供了一种为 “混排 TEX 文本和 S 编码” 生成文档的机制。单个的 Sweave 文档中既包含 TEX 文本又包含 S 编码，通过编译最终形成的文档包含：

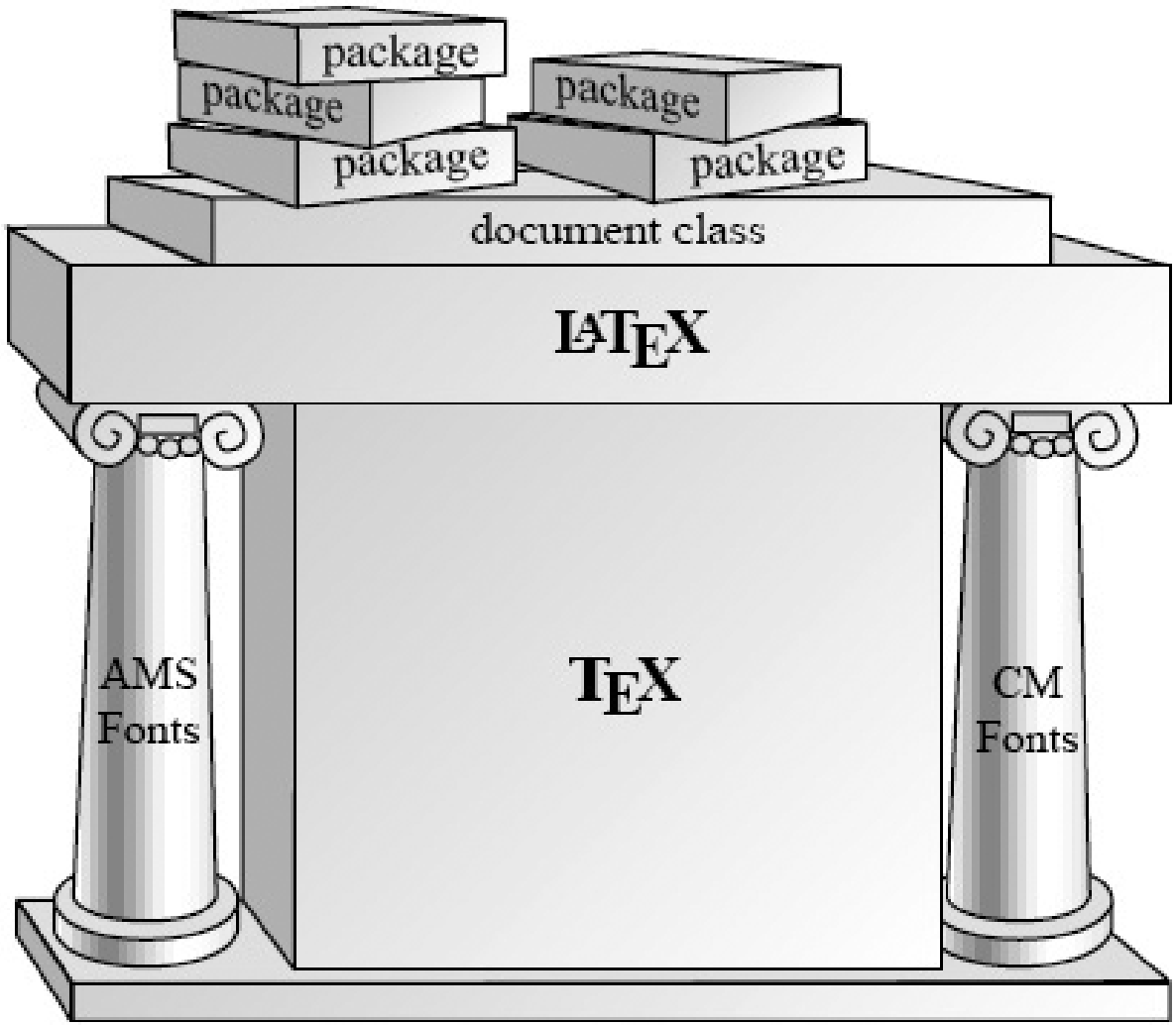
- TEX 文档的编译输出；
- S 编码和（或）；
- S 编码的代码输出（文本、图形）。

它的文档形成过程：

$$\text{Sweave 文档} \xrightarrow{\text{Sweave(in R)}} \text{TEX文档} \xrightarrow[\text{dvipdfmx or pdflatex}]{\text{LATEX}} \text{最终 pdf 文档}$$

什么是 LaTeX?

- 1977 年 5 月，Donald E. Knuth 最初用于 The Art of Computer Programming 的排版
- TeX 的版本号无限趋近于 π，目前为 3.141592
- 后来 Leslie Lamport 开发了基于 TeX的宏集 LaTeX


package
package
package
package
package
document class
LaTeX
AMS
Fonts
TeX
CM
Fonts

LaTeX 的优势

- 高质量的输出
- 超常的稳定性
- TeX 是可编程的
- 良好的通用性
- ……

一些花絮

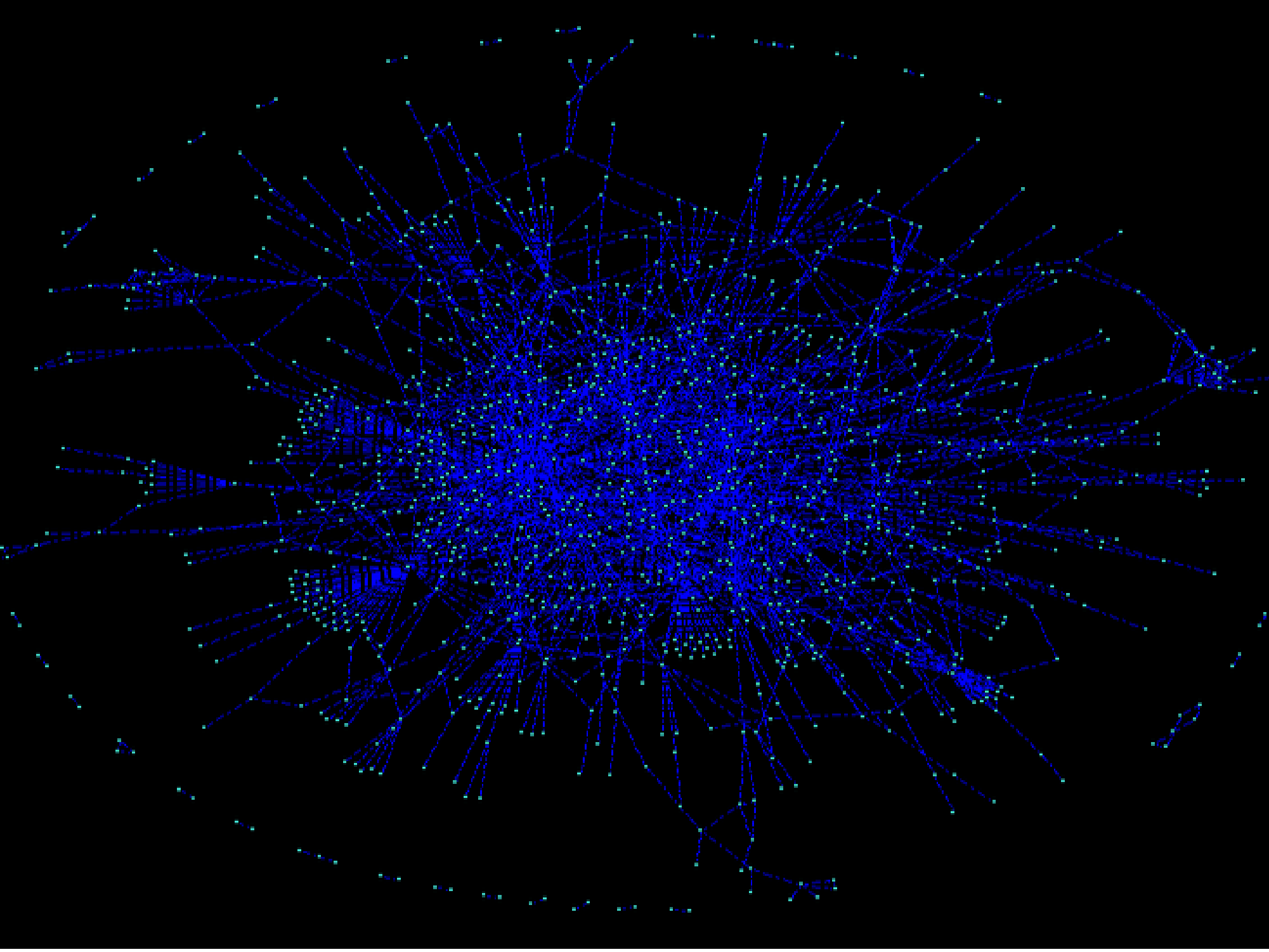
- R 的大部分帮助文档都是由 LaTeX 写的
- Comprehensive R(TeX) Archive Networks
- CTeX 的官网同样是 R 的中国镜像之一

```
\documentclass[a4paper,11pt]{article}
\usepackage{ctex}
\usepackage{verbatim}
\setlength{\textheight}{255mm}
\setlength{\topmargin}{-10mm} \setlength{\textwidth}{16.5cm}
\setlength{\oddsidemargin}{-6mm} \setlength{\columnseprule}{.1pt}
\setlength{\columnsep}{20pt}


\title{使用 Sweave 生成的自动化报告}
\author{第二届中国 R 语言会议\\
刘思吉吉}
\date{2009年12月5日}
\SweaveOpts{echo=FALSE}
\begin{document}
\maketitle
\thispagestyle{empty}

\section{前言}
使用 Sweave 可以很容易地将 \LaTeX{} 同 R 的代码混排文档转化为可编译的 \LaTeX{} 文档。

\section{一般性说明}
在这种混排的文档里，基本结构仍然是 \LaTeX{} 形式的，唯一的区别是，R 代码需要放置在以 $<<>>=$ 为开头，$@$为结尾的段落里面。开头部分有两个常用的参数：echo和fig,使用逻辑值分别表示是否将 R 代码输入作为 \LaTeX{} 文本输出；是否在 \LaTeX{} 文档中绘制图形。这篇文档只需要在 R 中编译一遍，即可形成\LaTeX{}需要的输出（文件）。

\section{t检验的例子}
下面是一个配对 t 检验的一个例子:
<<echo = TRUE>>=
m <- t.test(extra ~ group, data = sleep, paired = TRUE)
print(m)
@
R 在计算过程中生成的的中间结果很容易插入到标准文档，比如\texttt{sleep}数据的双样本的配对t检验结果中的$p$-value是\Sexpr{format.pval(m$p.value)};
或者是直接运算
<<echo=TRUE,results=hide>>=
choose(49,6)
@
美国威力球（类似于福彩双色球）的理论组合数等于\Sexpr{choose(49,6)}。通过这种方法处理‘‘有大量计算’’的文档，比 word 不知方便多少倍。

\section{插入图型}
使用 Sweave 还可以将 R 生成的图形加入到 \LaTeX{} 文档中，而不必事先做出 \LaTeX{} 需要的图形文件\footnote{Sweave会自动生成 ps 和 pdf 图形}。

下图是使用pdf绘图，100个随机数的散点图:

\begin{figure}[h]
  \centering
<< fig = TRUE, echo = FALSE,width = 12,height = 6 >>=
x <- rnorm(100)
par(mar = c(2,2,1,1))
plot(x,main = '',xlab = '',ylab = '')
@
\caption{The cats data from package MASS.}
\label{fig:cats}
\end{figure}

\section{绘制表格}
绘制表格的例子，这里调用了xtable包，使用xtable函数，通过函数里的参数对表格进行诸如小数点位数、标题的设置:

<<results=tex>>=
library(xtable)
xtable(WorldPhones,digits = 0, caption='The World Telephones')
@

\end{document}
```

- 阿基米德说：给我一个支点，我就能撬动地球！（棍子无限长）

- 阿基米德说：给我一个支点，我就能撬动地球！（棍子无限长）
- 而我也说：给我一个 R，我也能撬动地球！（包无限扩展）

寄语



- Email：sunbjt<at>gmail.com
- Blog：`http://www.bjt.name`

Jump to first slide