



**Data Mining with R/ORE**  
**Minming Duan**

# iTech Solution Profile

## Agenda

- 1** R/ORE Overview
- 2** XML output generation using SQL
- 3** Integration with IBP and BIEE
- 4** Oracle R for Hadoop Connector
- 5** R vs. SPSS
- 6** FAQ

# Why analysts use R

- R is a statistics language similar to Base SAS or SPSS statistics.
- R environment is...
  - • Powerful
  - • Extensible
  - • Graphical
  - • Extensive statistics
  - • OOTB functionality with many 'knobs' but smart defaults
  - • Ease of installation and use
  - • **Free**

# Limitations of R

R is a client and server bundled together as 1 executable - like Excel

- Single user tool
- Not multi-threaded
- Cannot leverage CPU capacity even on a user's laptop/desktop

R requires data it operates on to be first loaded into memory

- Loading data may not be a limitation given RAM available on laptops/desktops
- R's *call by value semantics* means as data flows into functions, for each function invocation, many copies of the data are made
- As a result you quickly run into memory limits

# Why should you be interested in R?

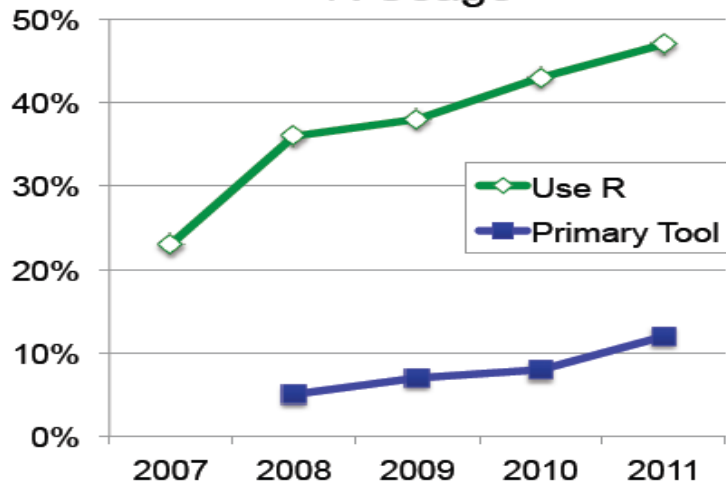
- Emerging trends
  - It's the next "big thing" in advanced analytics
  - Colleges and universities use R for statistics classes  
(replacing more traditional software tools)
  - Advanced Analytics as a critical differentiator of the DWH technology stack
- Augment Oracle deployments
  - Enhance results with powerful graphics
  - Integrate R results and graphics with BI Publisher documents and OBIEE dashboards
- A scalable R via Oracle R Enterprise
  - Leverage Oracle-engineered solutions
  - A viable alternative to SAS/SPSS

# Rexer Analytics Survey 2011

## The Popularity of R Software is Growing Fast

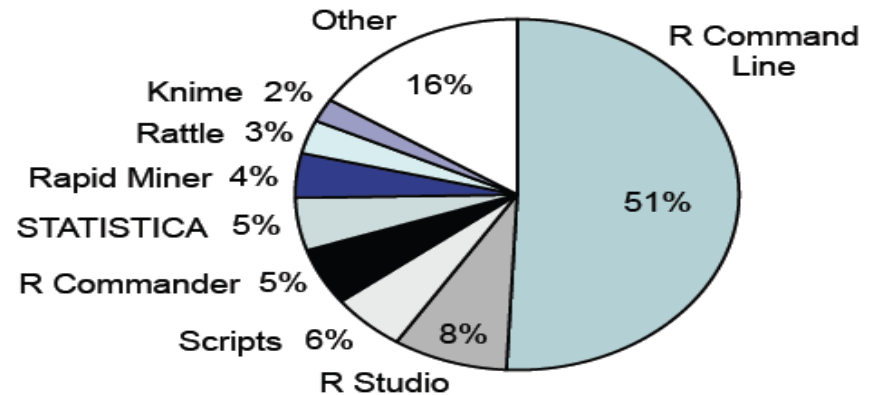
- The proportion of data miners using R is rapidly growing!
  - R is also the #1 most used data mining tool (in both 2010 & 2011). Up from #5 in 2007.
- An increasing number of data miners consider R their primary tool.
  - R is now #2 in primary tool rankings. Up from #7 in 2008.
- Half of R data miners use the command line interface. Among the rest, R Studio, scripts, R Commander, and STATISTICA are popular interfaces.

R Usage



Vendors were excluded from these analyses.

R Interface



Question: If you use the R software package, what is your primary interface to R?

# Default R GUI

The screenshot displays the R GUI interface. The R Console window on the left shows the following text:

```
R Console
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Error: package 'ROracle' is not installed for 'arch=x64'
R> library(ggplot2)
Loading required package: reshape
Loading required package: plyr

Attaching package: 'reshape'

The following object(s) are masked from 'package:plyr':

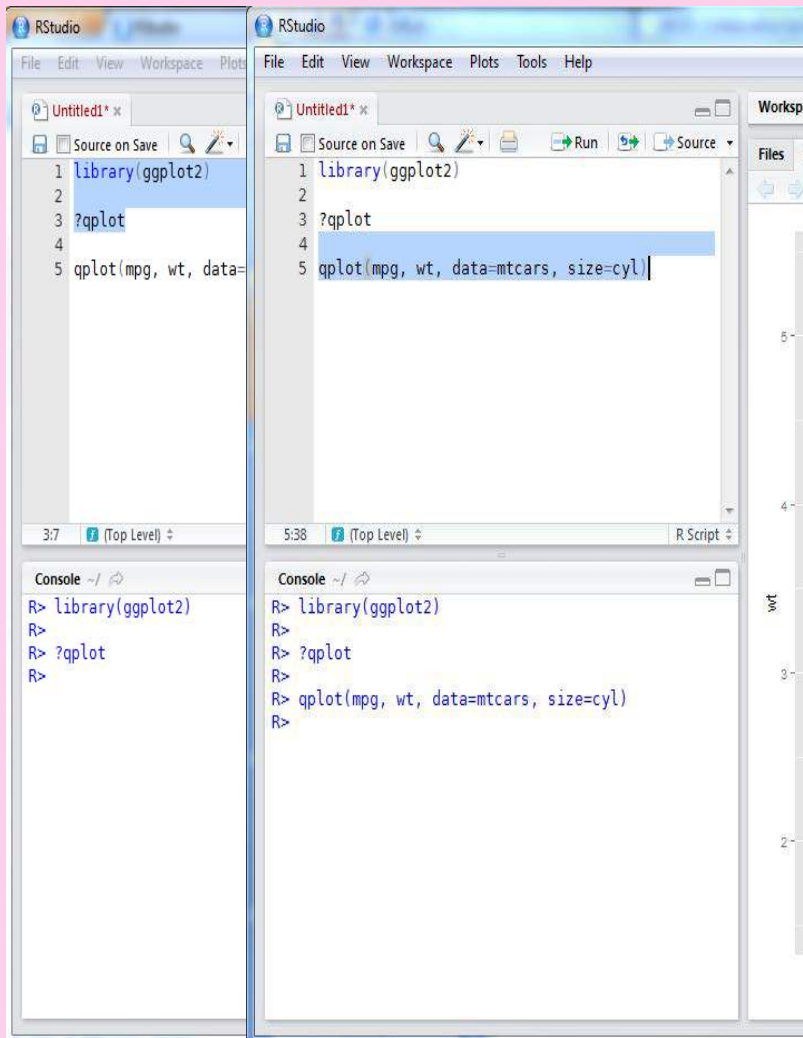
  rename, round_any

Loading required package: grid
Loading required package: proto
R> ?qplot
starting httpd help server ... done
R> qplot(mpg, wt, data=mtcars, facets=vs ~ am)
R> qplot(mpg, wt, data=mtcars, size=cyl)
R> |
```

The R Graphics window on the right, titled "R Graphics: Device 2 (ACTIVE)", displays a scatter plot of weight (wt) versus miles per gallon (mpg) for the mtcars dataset. The plot uses the size of the points to represent the number of cylinders (cyl). The x-axis (mpg) ranges from approximately 10 to 35, and the y-axis (wt) ranges from 2 to 5.5. A legend on the right indicates the mapping of point sizes to cylinder counts: 4 (smallest dot), 5 (small dot), 6 (medium dot), 7 (large dot), and 8 (largest dot).

mpg	wt	cyl
10.4	5.25	8
10.4	5.42	8
15.2	5.25	8
13.3	3.86	6
14.7	3.52	6
15.2	3.59	6
15.2	3.44	6
15.2	3.57	6
15.2	3.51	6
16.4	3.17	6
17.0	4.08	6
17.8	3.76	6
18.7	3.44	6
18.7	3.57	6
18.7	3.51	6
19.2	3.86	6
20.9	2.8	6
21.5	2.63	6
21.5	2.87	6
21.5	2.78	6
22.8	3.22	6
23.4	2.46	4
24.4	3.17	4
26.0	2.2	4
27.0	1.94	4
30.4	1.51	4
30.4	1.61	4
32.4	1.83	4

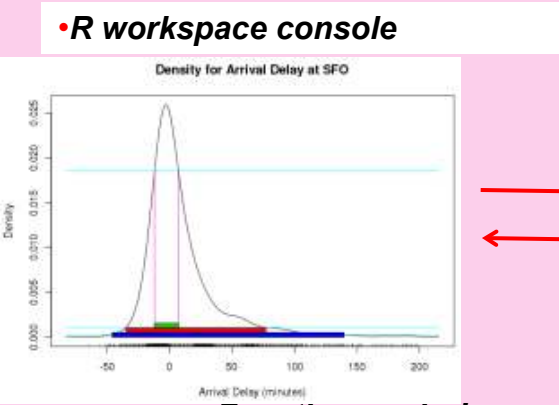
# RStudio – Third Party, Open Source IDE



Which R interfaces do you use frequently?	
built-in R console (225)	40%
RStudio (135)	24%
Eclipse with StatET (90)	16%
RapidMiner R extension (80)	14.2%
Tinn-R (62)	11%
ESS (Emacs Speaks Statistics) (59)	10.5%
Rattle GUI (53)	9.4%
R Commander (43)	7.7%
Revolution Analytics (31)	5.5%
RKward (22)	3.9%
JGR (Java Gui for R) (21)	3.7%
RExcel (18)	3.2%
R via a data mining tool plugin (12)	2.1%
Red-R (8)	1.4%
SciViews-R (6)	1.1%
Other (44)	7.8%

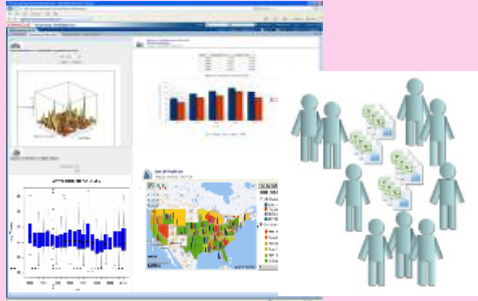
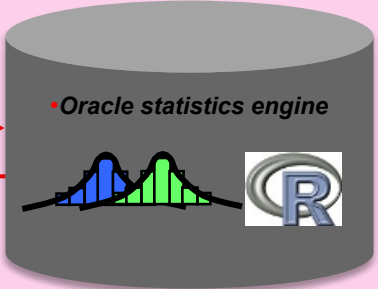


# Oracle R Enterprise



**•Function push-down**  
– data transformation & statistics

**ORACLE®**



**•No changes to the user experience**



**•Scale to large data sets**



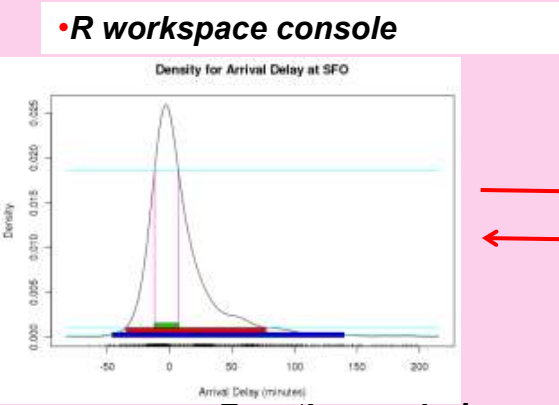
**•Embed in operational systems**

**•Development**

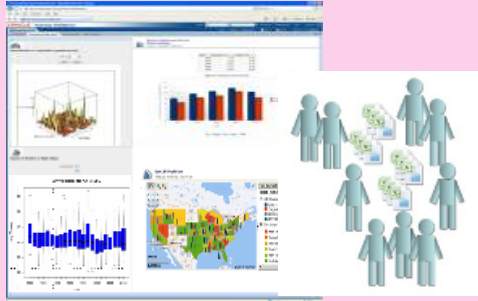
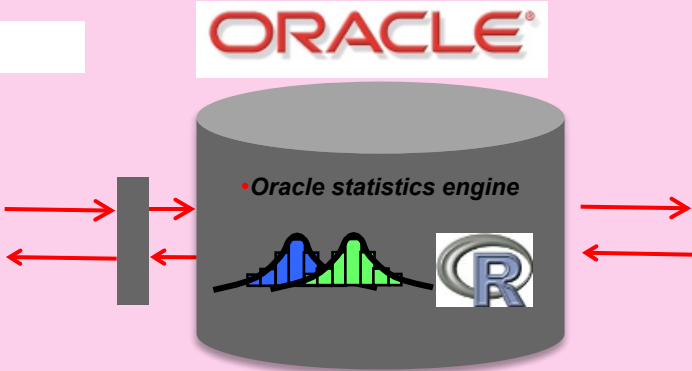
**•Production**

**•Consumption**

# Oracle R Enterprise



**•Function push-down**  
– data transformation & statistics



**•Transparently leverage Hadoop for High Performance Analytics to Oracle Big Data Appliance (part of Big Data Connectors software suite)**

# Oracle R Enterprise – Key messages

- ***Most integrated and complete suite of Enterprise Advanced Analytics software available in the market today***

- ***Substantial leap forward from incumbent platforms***
  - ***Data volume – using SQL and existing DB functionality***
  - ***Data Heterogeneity – Oracle DB + BDA***
  - ***Breadth of Analytics – Oracle DB + R packages***
  - ***Breadth of User Types – R+SQL+BI report developers, DBAs***

- ***Enables enterprise-wide consumption of advanced analytics models via integration with Oracle Exalytics***

# iTech Solution Profile

## Agenda

- 1 R/ORE Overview
- 2 XML output generation using SQL
- 3 Integration with IBP and BIEE
- 4 Oracle R for Hadoop Connector
- 5 R vs. SPSS
- 6 FAQ

# iTech Solution Profile

## Agenda

- 1 R/ORE Overview
- 2 XML output generation using SQL
- 3 Integration with IBP and BIEE
- 4 Oracle R for Hadoop Connector
- 5 R vs. SPSS
- 6 FAQ

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/496033241111010135>