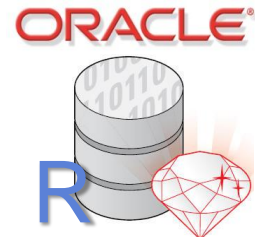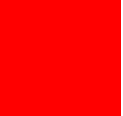# Oracle Advanced Analytics
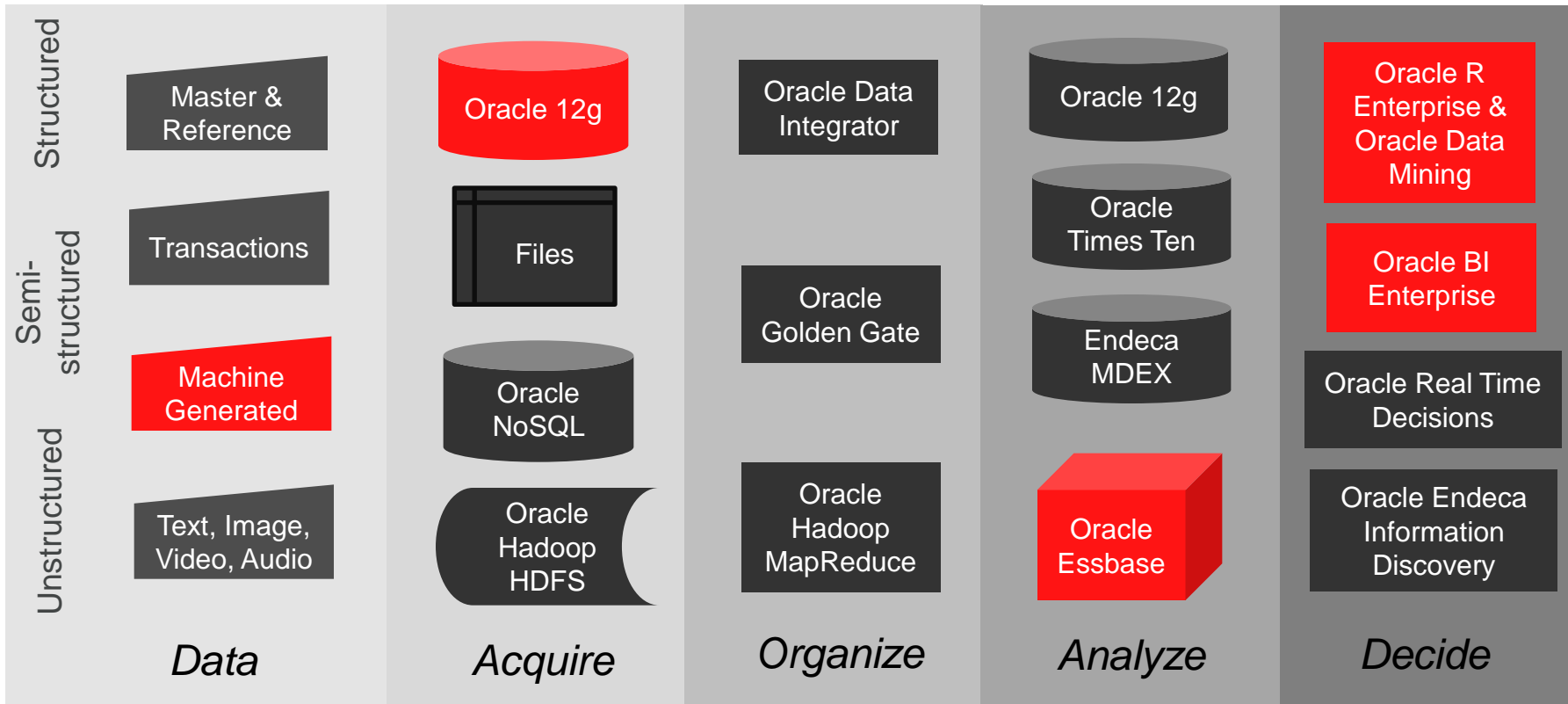**Oracle R Enterprise & Oracle Data Mining**

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.
The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

ORACLE®

USE CASE 8: ELECTRICAL CONSUMPTION

- Short Description :
  - CERN ENS (Electrical Network Supervisor) system archives values from measurements
  - Around 12000 every 10 minutes
  - The storage is managed by an external company on an instance of Oracle relational database.
  - Nevertheless, the DB instance is installed on one dedicated server connected to CERN Technical Network

- Issues:
  - Make data aggregation (by time and by families of measurements)
  - Electricity load forecast by system (families of measurements) using the historical data
  - Make the result obtained accessible to CERN users from General Purpose Network

ORACLE

Oracle Technology mapped to Analytics Landscape

	Data	Acquire	Organize	Analyze	Decide
Structured	Master & Reference	Oracle 12g	Oracle Data Integrator	Oracle 12g	Oracle R Enterprise & Oracle Data Mining
Semi-structured	Transactions	Files	Oracle Golden Gate	Oracle Times Ten	Oracle BI Enterprise
	Machine Generated	Oracle NoSQL		Endeca MDEX	Oracle Real Time Decisions
Unstructured	Text, Image, Video, Audio	Oracle Hadoop HDFS	Oracle Hadoop MapReduce	Oracle Essbase	Oracle Endeca Information Discovery

Data   *Acquire*   *Organize*   *Analyze*   *Decide*

Oracle Advanced Analytics Option—Agenda

Extending the Database into a Comprehensive Advanced Analytics Platform

- Oracle Data Mining
  - SQL & PL/SQL focused in-database data mining and predictive analytics

- Oracle R Enterprise
  - Integrates Open Source R with the Oracle Database

Oracle Advanced Analytics Option

Transforming the Database into a Comprehensive Advanced Analytics Platform

- Oracle Advanced Analytics Option enables companies to *"bring the algorithms to the data"* vs. extracting the data to specialized and expensive dedicated statistical and data mining servers

- Oracle Advanced Analytics Option includes:
  - Oracle Data Mining
    - SQL & PL/SQL focused in-database data mining and predictive analytics
  - Oracle R Enterprise
    - Integrates the Open-Source Statistical Environment R with the Oracle Database

- Data movement is eliminated or dramatically reduced while analytical and compute intensive operations are performed inside the database

Oracle Data Mining

Building Predictive Analytics Applications
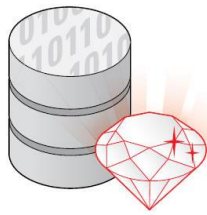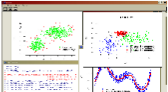
- Oracle Data Mining provides 12 powerful in-database data mining algorithms for big data analytics as a native feature of the database
  - Designed for or big data problems involving discovering patterns and relationships in large amounts of data and oftentimes making predictions based on those patterns, Oracle Data Mining allows data analysts and data miners to mine star schemas, transactional data and unstructured data stored inside the database, build predictive models and apply them to data inside the database--all without moving data.

- Developers can use the Oracle Data Miner extension to SQL Developer to develop, build, evaluate, share and automate analytical workflows to solve important data driven business problems.

- Developers can use the SQL APIs and PL/SQL to build applications to automate knowledge discovery
  - The Oracle Data Miner GUI generates SQL code that application developers can use to develop and deploy SQL and PL/SQL based automated predictive analytics applications that run natively inside the Oracle Database.

What is Data Mining?

- Automatically finds hidden patterns, discover new insights, and make predictions

- Data Mining can provide valuable results:
  - Predict customer behavior *(Classification)*
  - Predict or estimate a value *(Regression)*
  - Segment a population *(Clustering)*
  - Identify factors more associated with a business problem *(Attribute Importance)*
  - Find profiles of targeted people or items *(Decision Trees)*
  - Determine important relationships and "market baskets" within the population *(Associations)*
  - Find fraudulent or "rare events" *(Anomaly Detection)*
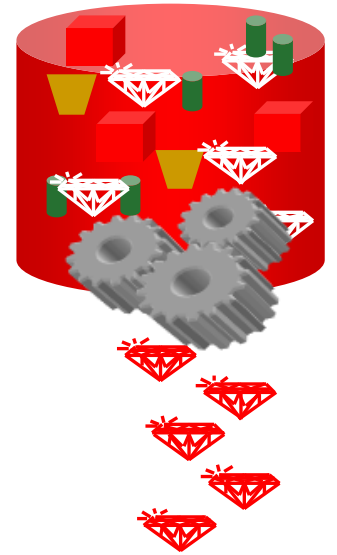
SQL Developer 3.0/Oracle Data Miner 11g Release 2 GUI

- Graphical User Interface for data analyst
- SQL Developer Extension (OTN download)
- Explore data—discover new insights
- Build and evaluate data mining models
- Apply predictive models
- Share analytical workflows
- Deploy SQL Apply code/scripts

ORACLE

Oracle Data Miner Nodes (Partial List)

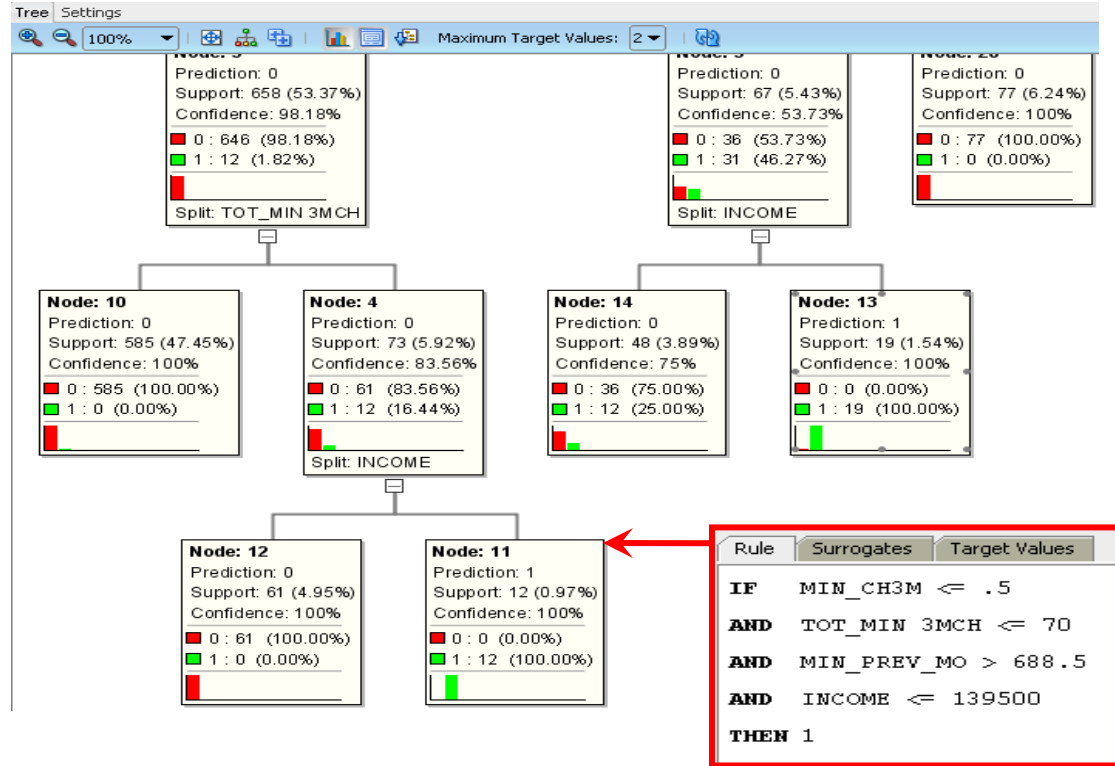

Tables and Views
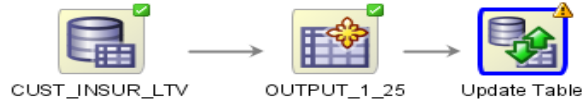
Transformations
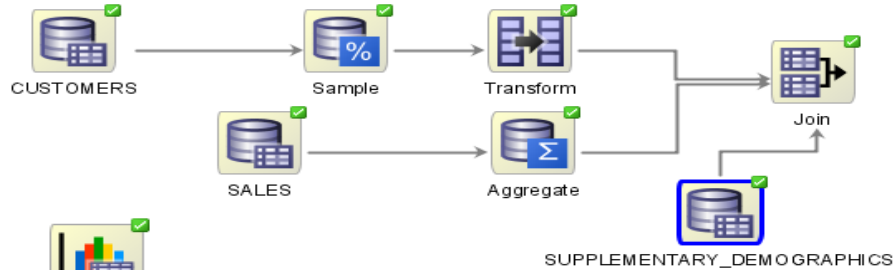
Explore Data

Modeling
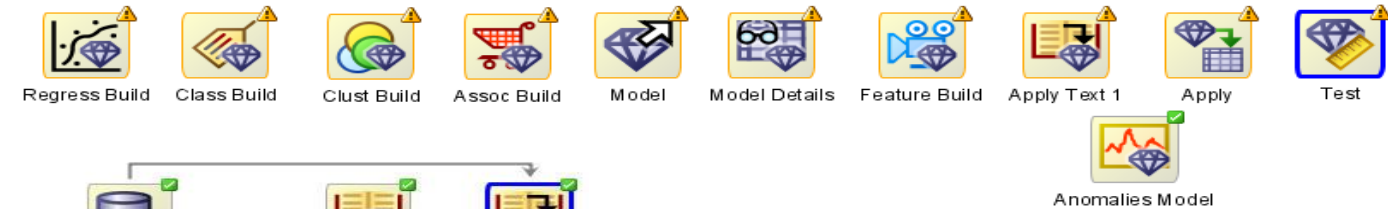
Text
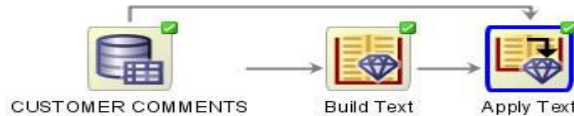
ORACLE

Oracle Data Miner 11g Release 2 GUI
Churn Demo—Simple Conceptual Workflow
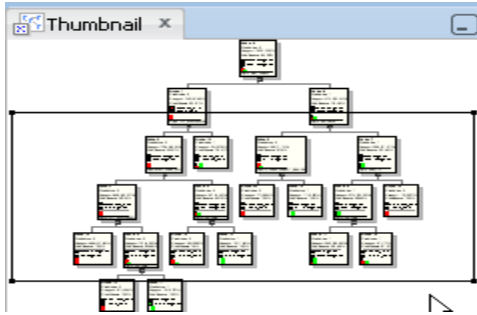
Churn models to product and "profile" likely churners

| Tree | Settings |

Maximum Target Values: 2

Node: 3
Prediction: 0
Support: 658 (53.37%)
Confidence: 98.18%
- 0 : 646 (98.18%)
- 1 : 12 (1.82%)
Split: TOT_MIN 3MCH

Node: 5
Prediction: 0
Support: 67 (5.43%)
Confidence: 53.73%
- 0 : 36 (53.73%)
- 1 : 31 (46.27%)
Split: INCOME

Node: 20
Prediction: 0
Support: 77 (6.24%)
Confidence: 100%
- 0 : 77 (100.00%)
- 1 : 0 (0.00%)

Node: 10
Prediction: 0
Support: 585 (47.45%)
Confidence: 100%
- 0 : 585 (100.00%)
- 1 : 0 (0.00%)

Node: 4
Prediction: 0
Support: 73 (5.92%)
Confidence: 83.56%
- 0 : 61 (83.56%)
- 1 : 12 (16.44%)
Split: INCOME

Node: 14
Prediction: 0
Support: 48 (3.89%)
Confidence: 75%
- 0 : 36 (75.00%)
- 1 : 12 (25.00%)

Node: 13
Prediction: 1
Support: 19 (1.54%)
Confidence: 100%
- 0 : 0 (0.00%)
- 1 : 19 (100.00%)

Node: 12
Prediction: 0
Support: 61 (4.95%)
Confidence: 100%
- 0 : 61 (100.00%)
- 1 : 0 (0.00%)

Node: 11
Prediction: 1
Support: 12 (0.97%)
Confidence: 100%
- 0 : 0 (0.00%)
- 1 : 12 (100.00%)

| Rule | Surrogates | Target Values |

IF   MIN_CH3M <= .5
AND TOT_MIN 3MCH <= 70
AND MIN_PREV_MO > 688.5
AND INCOME <= 139500
THEN 1

ORACLE

Oracle Data Miner 11g Release 2 GUI

Simple Conceptual Workflow
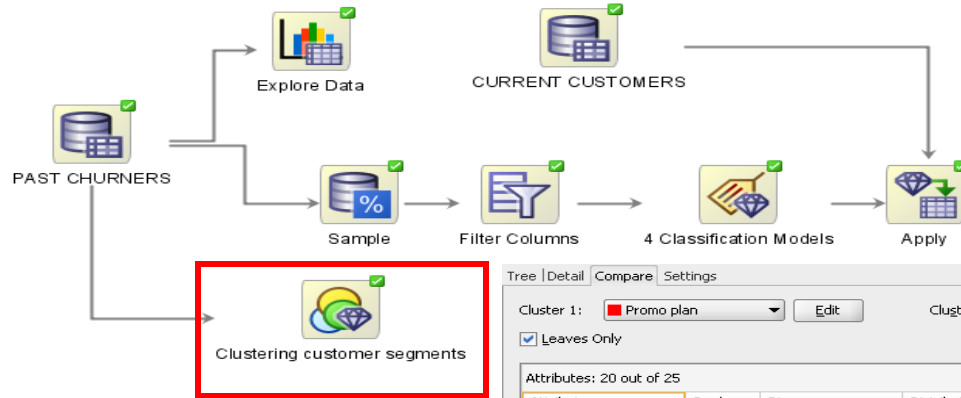
Clustering analysis to discover customer segments based on behavior, demograhics, plans, equipment, etc.

ORACLE®

Fraud Prediction Demo

```
drop table CLAIMS_SET;
exec dbms_data_mining.drop_model('CLAIMSMODEL');
create table CLAIMS_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into CLAIMS_SET values
    ('ALGO_NAME','ALGO_SUPPORT_VECTOR_MACHINES');
insert into CLAIMS_SET values ('PREP_AUTO','ON');
commit;

begin
dbms_data_mining.create_model('CLAIMSMODEL', 'CLASSIFICATION',
 'CLAIMS', 'POLICYNUMBER', null, 'CLAIMS_SET');
end;
/

-- Top 5 most suspicious fraud policy holder claims
select * from
(select POLICYNUMBER, round(prob_fraud*100,2) percent_fraud,
    rank() over (order by prob_fraud desc) rnk from
(select POLICYNUMBER, prediction_probability(CLAIMSMODEL, '0' using *) prob_fraud
from CLAIMS
where PASTNUMBEROFCLAIMS in ('2to4', 'morethan4')))
where rnk <= 5
order by percent_fraud desc;
```

POLICYNUMBER	PERCENT_FRAUD	RNK
6532	64.78	1
2749	64.17	2
3440	63.22	3
654	63.1	4
12650	62.36	5

Automated Monthly "Application"! *Just add:*
    Create
    View CLAIMS2_30
    As
    Select * from CLAIMS2
    Where mydate > SYSDATE − 30

ORACLE®

ORACLE®

Exadata + Data Mining 11g Release 2

"DM Scoring" Pushed to Storage!

Faster

- In 11g Release 2, SQL predicates and Oracle Data Mining models are <u>pushed to storage level for execution</u>
  For example, find the US customers likely to churn:

```
select cust_id
from customers
where region = 'US'
and prediction_probability(churnmod,'Y' using *) > 0.8;
```

Scoring function executed in Exadata

ORACLE

Oracle Communications Industry Data Model
Better Information for OBIEE Dashboards

Oracle Data Mining identifies key contributors to customer churn
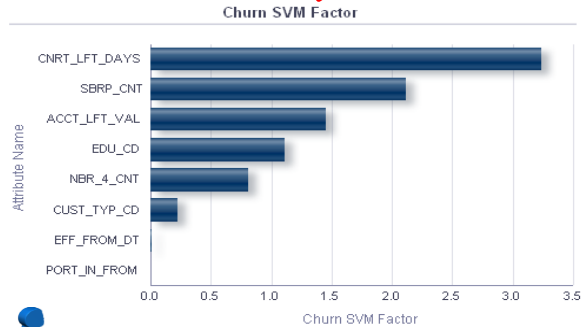
Oracle Data Mining discovers different profiles of churning customers and their profile, both critical to developing a proactive response to reduce churn.

Oracle Communications Industry Data Model Example
Better Information for OBIEE Dashboards

ODM's predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools

Oracle Data Mining Algorithms

Problem	Algorithm	Applicability
Classification	Logistic Regression (GLM) Decision Trees Naïve Bayes Support Vector Machine	Classical statistical technique Popular / Rules / transparency Embedded app Wide / narrow data / text
Regression	Multiple Regression (GLM) Support Vector Machine	Classical statistical technique Wide / narrow data / text
Anomaly Detection	One Class SVM	Lack examples of target field
Attribute Importance	Minimum Description Length (MDL)	Attribute reduction Identify useful data Reduce data noise
Association Rules	Apriori	Market basket analysis Link analysis
Clustering	Hierarchical K-Means Hierarchical O-Cluster	Product grouping Text mining Gene and protein analysis
Feature Extraction	Nonnegative Matrix Factorization	Text analysis Feature reduction

ORACLE

Learn More

Oracle Data Mining on OTN

Oracle Data Mining Blog

Oracle Data Mining

R Statistical Programming Language
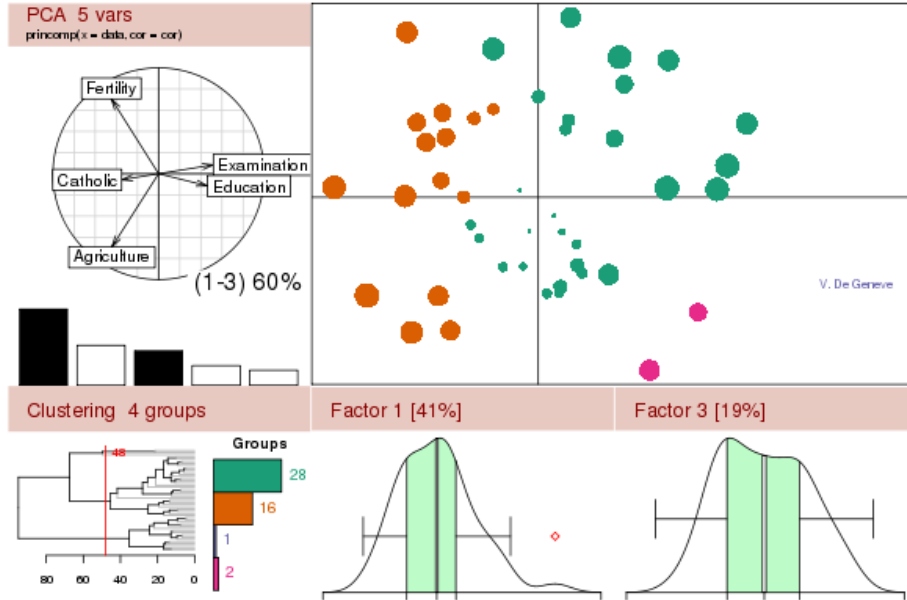
Open source language and environment

Used for statistical computing and graphics

Strength in easily producing publication-quality plots

Highly extensible with open source community R packages

Growing Popularity

The R Project for Statistical Computing
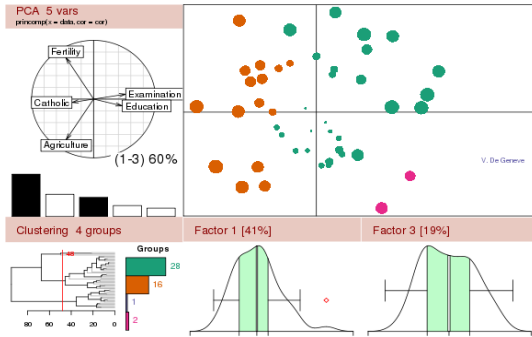
- R's rapid adoption over several years has earned its reputation as a new statistical software standard
  - Rival to SAS and SPSS
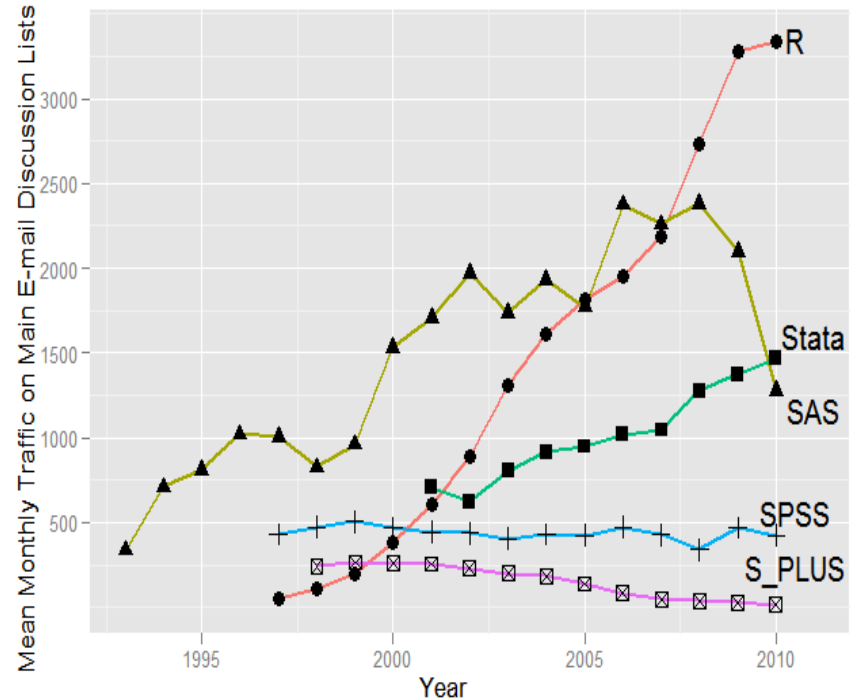
While it is difficult to calculate exactly how many people use R, those most familiar with the software estimate that close to 250,000 people work with it regularly.
"Data Analysts Captivated by R's Power", New York Times, Jan 6, 2009

http://www.r-project.org/

ORACLE

Typical R Approach
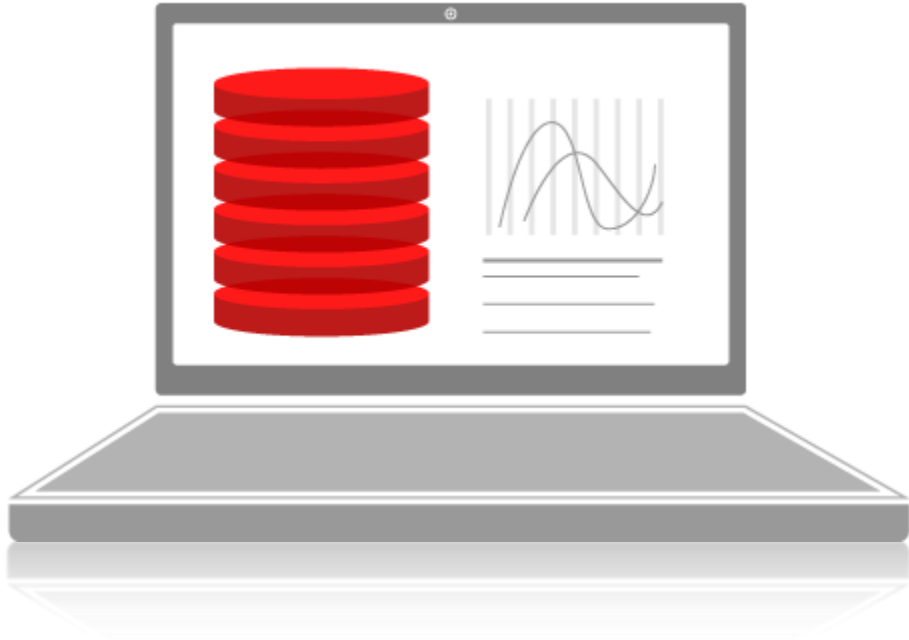
Statistical and advanced analyses are run and stored on the user's laptop

ORACLE®

What Are ⓡ 's Challenges?

1. **R is memory constrained**
   - R processing is single threaded - does not exploit available compute infrastructure
   - R lacks industrial strength for enterprise use cases

2. **R has lacked mindshare in Enterprise market**
   - R is still met with caution by the long established SAS and IBM/SPSS statistical community
     - However, major university (e.g. Yale) Statistics courses now taught in R
     - The FDA has recently shown indications for approval of new drugs for which the submission's data analysis was performed using R

ORACLE

Oracle R Enterprise Approach

Faster

Highly Secure

R
Open Source

Scalable
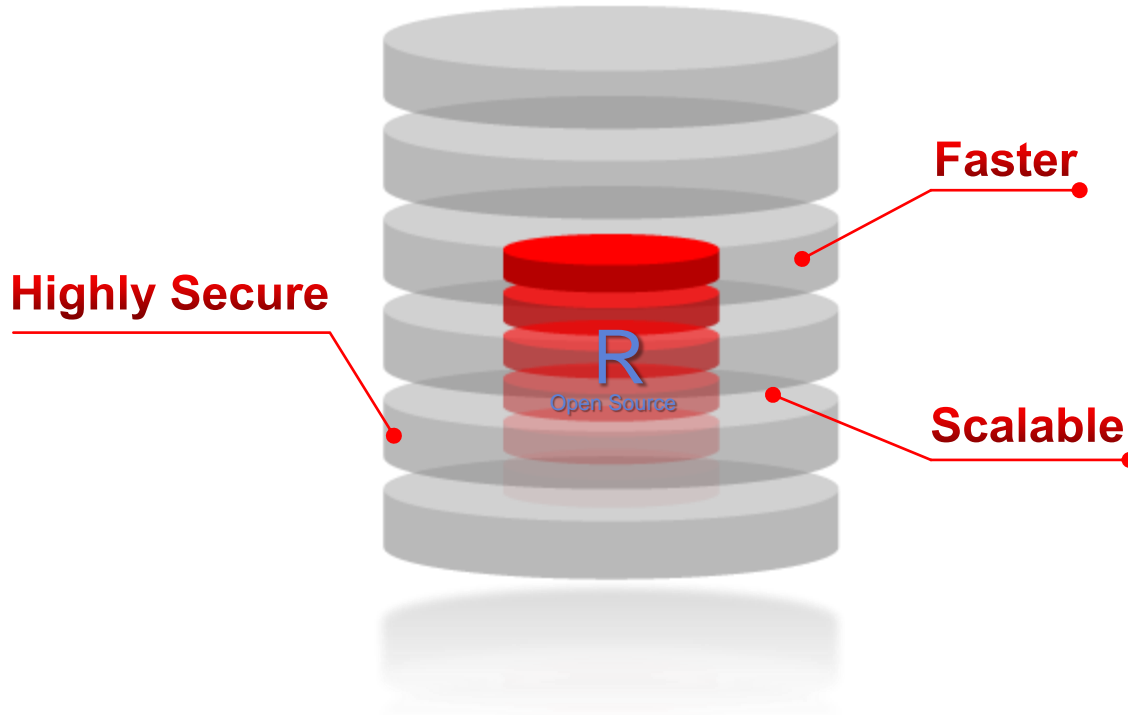
Data and statistical analysis are stored and run in-database

Same R user experience & same R clients

Embed in operational systems

Complements Oracle Data Mining

ORACLE®

What is ORACLE R Enterprise?

- Oracle R Enterprise brings R's statistical functionality closer to the Oracle Database

1. Eliminate R's memory constraint by enabling R to work directly/transparently on database objects
   - Allows R to run on very large data sets, tables, views

2. Architected for Enterprise production infrastructure
   - Automatically exploits database parallelism without requiring parallel R programming
   - Build and immediately deploy R scripts
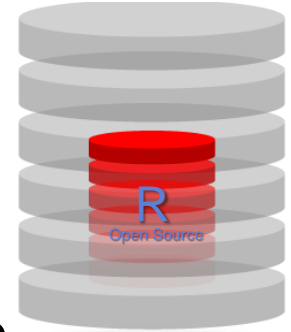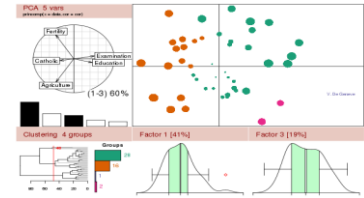
How Oracle R Enterprise Works
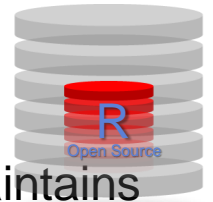
ORE Computation Engines

- Oracle R Enterprise eliminates data movement and duplication, maintains security and minimizes latency time from raw data to new information.
  - The database is always involved in serving up data to the R code.
  - Oracle R Enterprise runs in the Oracle Database

- Three ORE Computation Engines
  - Oracle R Enterprise provides three different interfaces between the open-source R engine and the Oracle database:

    1. Oracle R Enterprise (ORE) Transparency Layer
    2. Oracle Statistics Engine
    3. Embedded R

ORACLE

Oracle R Enterprise Compute Engines
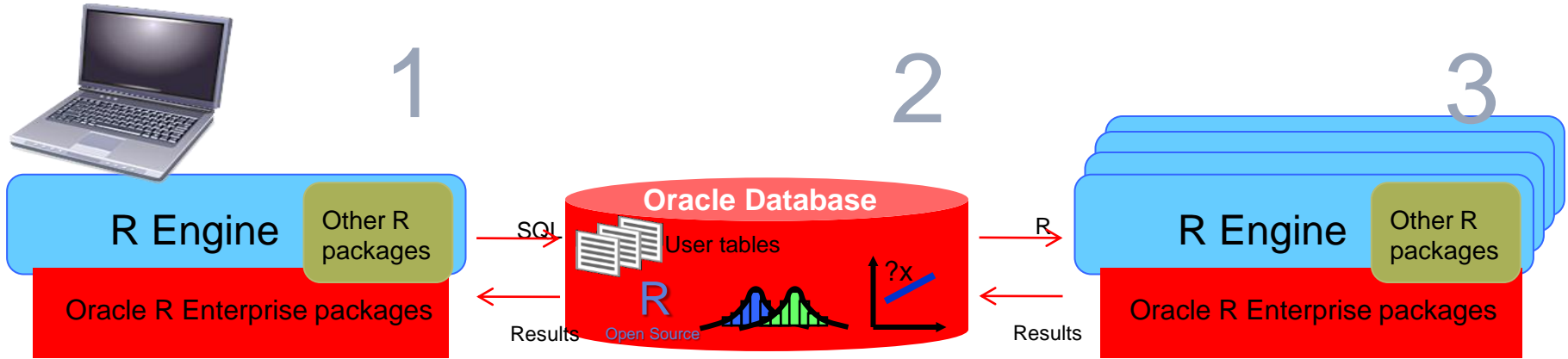
User R Engine on desktop

- R-SQL Transparency Framework intercepts R functions for scalable in-database execution
- Submit entire R scripts for execution by Oracle Database

Database Compute Engine

- Access tables, views, and external tables, as well as data through DB LINKS
- Leverage database SQL parallelism
- Leverage new and existing in-database statistical and data mining capabilities

R Engine(s) spawned by Oracle DB

- Database can spawn multiple R engines for database-managed parallelism
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables "lights-out" execution of R scripts

ORACLE

How Oracle R Enterprise Works

ORE Computation Engines

1. ## Oracle R Enterprise (ORE) Transparency Layer

   – Traps all R commands and scripts prior to execution and looks for opportunities to function ship them to the database for native execution

   – ORE transparency layer converts R commands/scripts into SQL equivalents and thereby leverages the database as a compute engine.
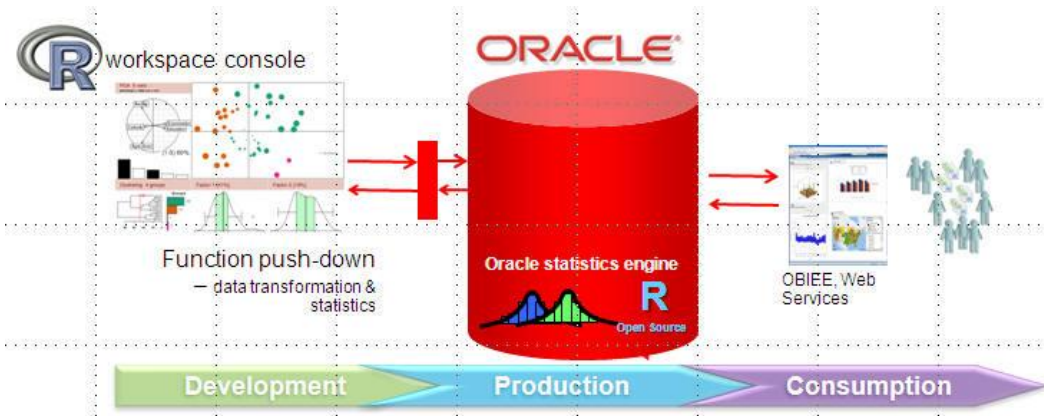
How Oracle R Enterprise Works

ORE Computation Engines
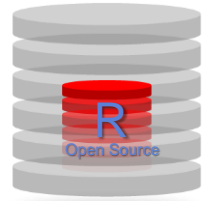
2. In-Database Statistics Engine

   – Significantly extends the Oracle Database's library of statistical functions and advanced analytical computations

   – Provides support for the complete R language and statistical functions found in Base R and selected R packages based on customer usage

      • Open source packages - written entirely in R language with only the functions for which we have implemented SQL counterparts - can be translated to execute in database.

   – Without anything visibly different to the R users, their R commands and scripts are oftentimes accelerated by a factor of 10-100x

**All Base R functions
R Multiple Regression**

…. Driven by customers

ORE Functions
• ORE SUMMARY
• ORE FREQUENCY
• ORE CORR
• ORE UNIVARITE
• ORE CROSSTAB
• ORE RANK
• ORE SORT
• …

ORACLE

How Oracle R Enterprise Works

ORE Computation Engines
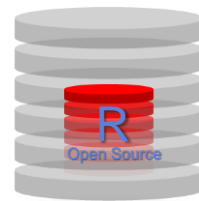
3. Embedded R Engine

- For R functions not able to be mapped to native in-database functions, Oracle R Enterprise makes "extproc" remote procedure calls to multiple R engines running on multiple database servers/nodes

- This Oracle R Enterprise embedded layer uses the database as a data provider providing data level parallelism to R code

ORACLE

Working with ORE Connections

```
ore.connect("RQUSER","SID","HOST","PASSWORD",1521)
```

Connect to a specific schema and database

One connection active at a time

```
ore.create( ONTIME_S, table = "NEW_ONTIME_S")
ore.create( ONTIME_S, view = "NEW_ONTIME_S_VIEW")
```

Create a database table from a data.frame, ore.frame. Create a view from an ore.frame.

```
ore.drop(table="NEW_ONTIME_S")
ore.drop(view="NEW_ONTIME_S_VIEW")
```

Drop table or view in database

```
v <- ore.push(c(1,2,3,4,5))
```

Store R object in database as temporary object, returns handle to object. Data frame, matrix, and vector to table, list/model/others to serialized object

```
ore.sync()
ore.sync("RQUSER")
ore.sync(table=c("ONTIME_S", "NARROW"))
ore.sync("RQUSER", table=c("ONTIME_S", "NARROW"))
```

Synchronize ORE proxy objects in R with tables/views available in database, on a per schema basis

```
ore.exists("ONTIME_S", "RQUSER")
```

Returns TRUE if named table or view exists in schema

ORACLE

Working with ORE Connections

```
ore.ls()
ore.ls("RQUSER")
ore.ls("RQUSER",all.names=TRUE)
ore.ls("RQUSER",all.names=TRUE, pattern= "NAR")


t <- ore.get("ONTIME_S","RQUSER")


ore.attach("RQUSER")
ore.attach("RQUSER", pos=2)



ore.detach("RQUSER")


ore.rm("DF1")
ore.rm(list("TABLE1","TABLE2"), "RQUSER")



ore.disconnect()


ore.exec("create table F2 as select * from ONTIME_S")
```

List the objects available in ORE environment mapped to database schema.

All.names=FALSE excludes names starting with a '.'

Obtain object to named table/view in schema.

Make database objects visible in R for named schema. Can place corresponding environment in specific position in env path.

Remove schema's environment from the object search path.

Remove table or view from schema's R environment.

Disconnect from the database. Clean up all associated R objects and temporary database schema objects

Execute SQL or PL/SQL without return value

ORACLE

Create database tables from data.frames

```
df <- data.frame(A=1:26, B=letters[1:26])

dim(df)

class(df)

ore.create(df,table="TEST_DF")

ore.ls(pattern="TEST_DF")

class(TEST_DF)

dim(TEST_DF)

head(TEST_DF)

ore.drop(table="TEST_DF ")
```

```
R> df <- data.frame(A=1:26, B=letters[1:26])
R> dim(df)
[1] 26  2
R> class(df)
[1] "data.frame"
R> ore.create(df,table="TEST_DF")
R> ore.ls(pattern="TEST_DF")
[1] "TEST_DF"
R> class(TEST_DF)
[1] "ore.frame"
attr(,"package")
[1] "OREbase"
R> dim(TEST_DF)
[1] 26  2
R> head(TEST_DF)
  A B
0 1 a
1 2 b
2 3 c
3 4 d
4 5 e
5 6 f
R> ore.drop(table="TEST_DF ")
```

R user on desktop

Client R Engine   Other R packages

Transparency Layer

Oracle R package

Oracle Database

User tables

Goal: Create a database table named "TEST_DF" from an R data.frame object.

Enables user to transition from laptop files to database schema

ORACLE
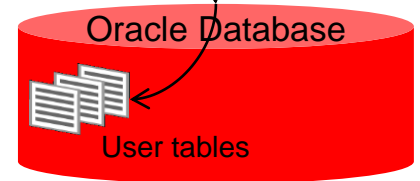
SQL and Parallel Execution

ORACLE®

Embedded Script Execution – SQL Interface

R Interface function	Purpose
rqEval()	Invoke stand-alone R script
rqTableEval()	Invoke R script with full table as input
rqRowEval()	Invoke R script on one row at a time, or multiple rows in chunks
rqGroupEval()	Invoke R script on data partitioned by grouping column
sys.rqScriptCreate	Create named R script
sys.rqScriptDrop	Drop named R script

ORE function	Input data	FUN.VALUE	Arguments	R Script	Special
rqEval()	Internally generated data	NULL (returns chunked blob)	NULL	String argument	Not applicable

```
begin
  sys.rqScriptCreate('Example1',
'function() {
   ID <- 1:10
   res <- data.frame(ID = ID, RES = ID / 100)
   res}');
end;
/
select *
  from table(rqEval(NULL,
      'select 1 id, 1 res from dual',
      'Example1'));
```

```
SQL> begin
  sys.rqScriptCreate('Example1',
'function() {
   ID <- 1:10
   res <- data.frame(ID = ID, RES = ID / 100)
   res}');
end;
/
select *
  from table(rqEval(NULL,
      'select 1 id, 1 res from dual',
      'Example1'));
  2    3    4    5    6    7    8
PL/SQL procedure successfully completed.

SQL>  2    3    4
      ID          RES
---------- ----------
        1         .01
        2         .02
        3         .03
        4         .04
        5         .05
        6         .06
        7         .07
        8         .08
        9         .09
       10          .1

10 rows selected.
```

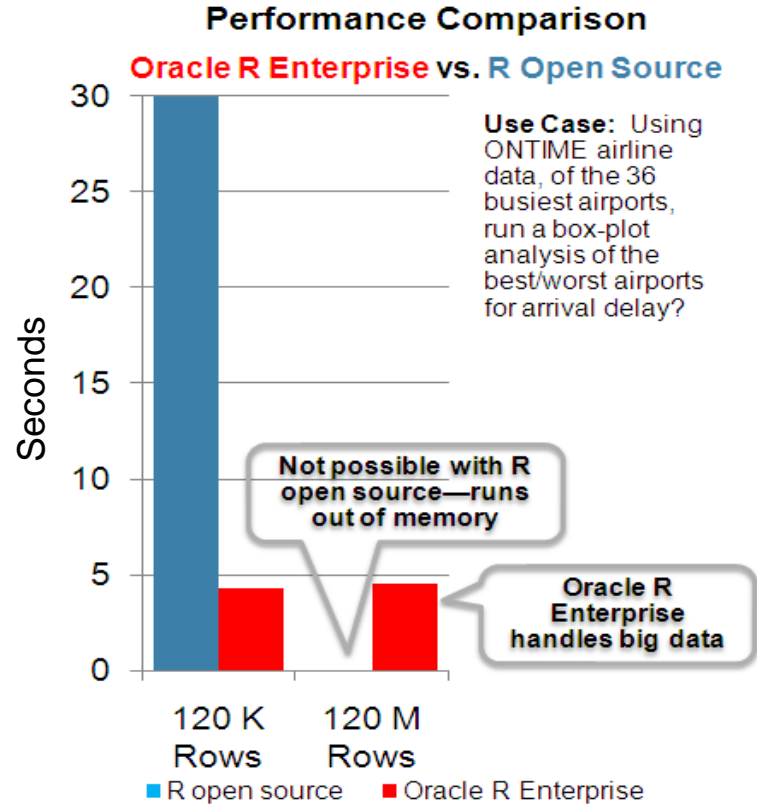- Execute R script with no parameters
- Specify output to return two numbers
  - id
  - res

Parallelism in the Transparency Layer

- Ideal for "bigger data",
- Operations performed in-database leverage database parallelism
- Database and table must be configured for parallelism as above
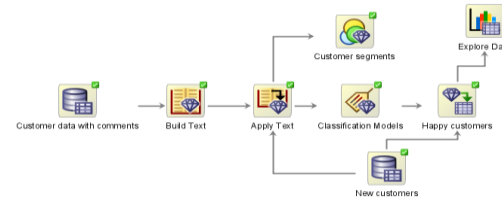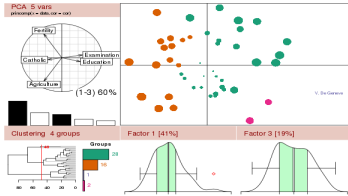
ORACLE

Architecture and Performance

- Transparently function-ships R constructs to database via R → SQL translation

- Performs data-heavy computations in database
  –R for summary analysis and graphics

- Transparent implementation enables using wide range of R "packages" from open source community

Performance Comparison
Oracle R Enterprise vs. R Open Source

Use Case: Using ONTIME airline data, of the 36 busiest airports, run a box-plot analysis of the best/worst airports for arrival delay?

Not possible with R open source—runs out of memory

Oracle R Enterprise handles big data

Seconds

120 K Rows   120 M Rows

■ R open source   ■ Oracle R Enterprise

ORACLE

Oracle In-Database Advanced Analytics

Comprehensive Advanced Analytics Platform

Oracle R Enterprise

- Popular open source statistical programming language & environment
- Integrated with database for scalability
- Wide range of statistical and advanced analytical functions
- R embedded in enterprise appls & OBIEE
- Exploratory data analysis
- Extensive graphics
- Open source R (CRAN) packages
- Integrated with Hadoop for HPC
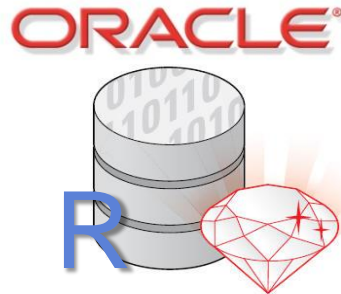
Oracle Data Mining

- Automated knowledge discovery inside the Database
- 12 in-database data mining algorithms
- Text mining
- Predictive analytics applications development environment
- Star schema and transactional data mining
- Exadata "scoring" of ODM models
- SQL Developer/Oracle Data Miner GUI

Statistics   Advanced Analytics   Data & Text Mining   Predictive Analytics

ORACLE®

ORACLE®