



ORACLE  
OPEN  
WORLD

SOFTWARE. HARDWARE. **COMPLETE.**

# ORACLE®

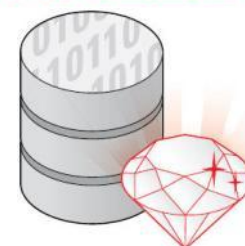
## ***Oracle Data Mining — In-Database Data Mining Made Easy!***

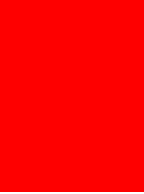
Charlie Berger  
Sr. Director Product Management, Data Mining and Advanced Analytics  
Oracle Corporation  
[charlie.berger@oracle.com](mailto:charlie.berger@oracle.com)  
[www.twitter.com/CharlieDataMine](http://www.twitter.com/CharlieDataMine)

*Copyright 2010 Oracle Corporation*



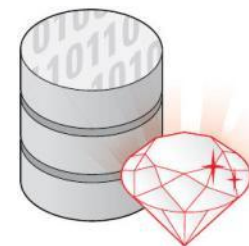
ORACLE®





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.

# Agenda



- Market Drivers
- Oracle Data Mining
- Exadata and Oracle Data Mining
- Oracle Data Miner 11g Release 2 New GUI
- Oracle Statistical Functions
- Ability to Import 3<sup>rd</sup> Party e.g. SAS models
- Applications Powered by Oracle Data Mining
- Getting Started with ODM

A man in a dark suit, light blue shirt, and striped tie is sitting in a black leather office chair. He is gesturing with his right hand, palm facing up. Behind him is a large, silver Oracle server rack with a perforated front panel. The rack has various labels and controls, including 'TAPE', 'STANDBY', and 'XSCF'. The background is a blurred office setting with large windows.

# Market Drivers

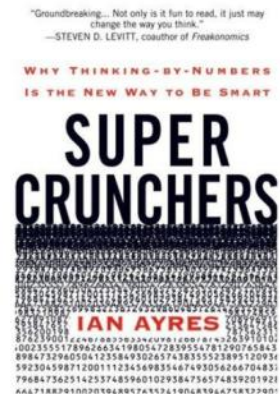
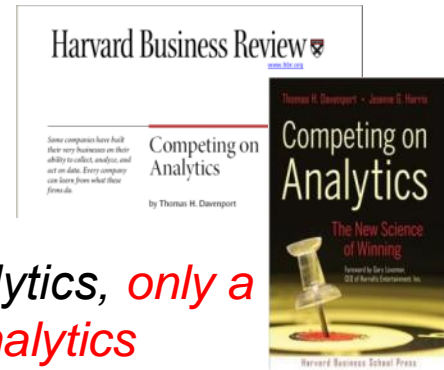
SOFTWARE.  
HARDWARE.  
**COMPLETE.**

ORACLE®

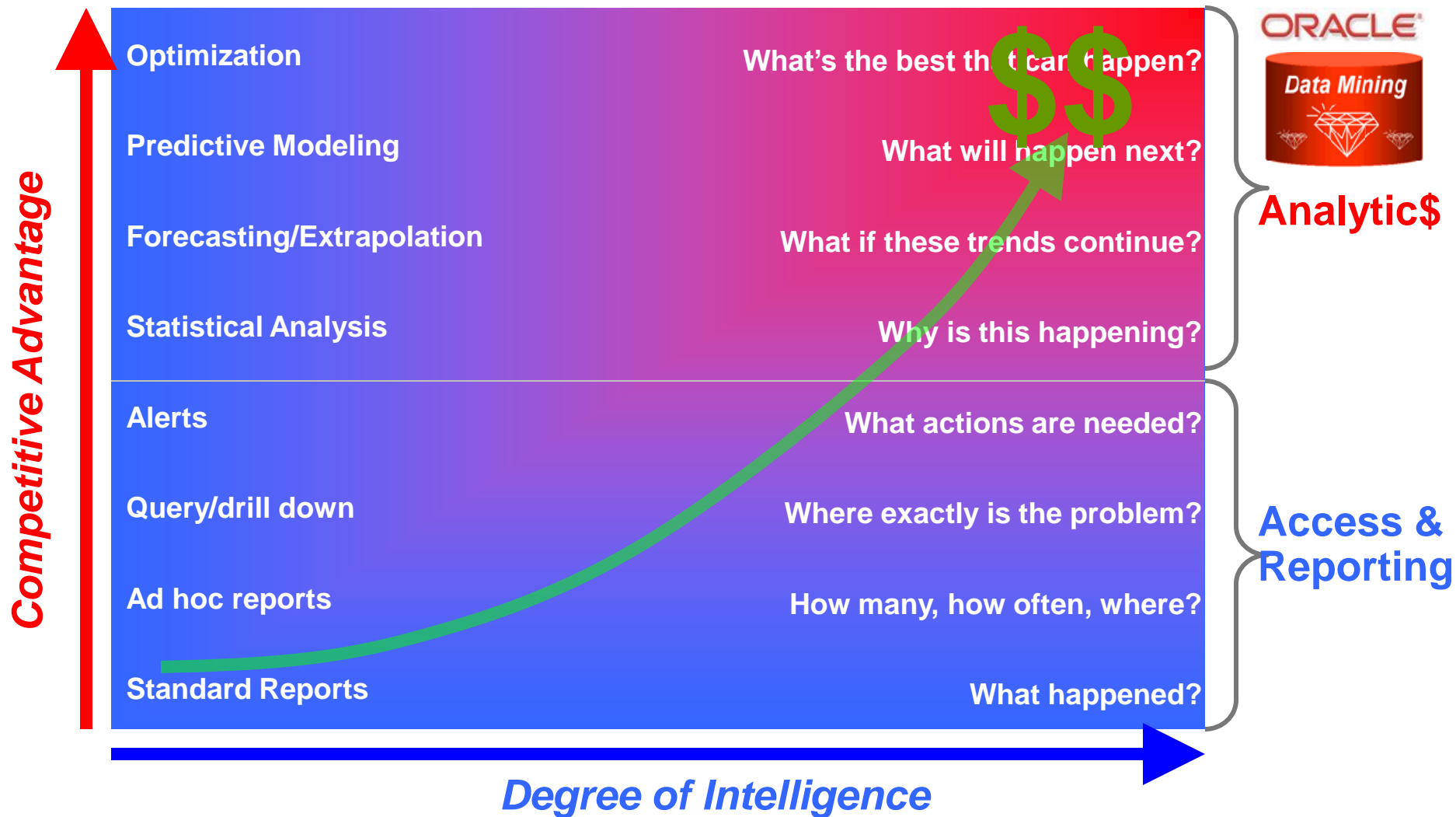


# Analytics: Strategic and Mission Critical

- *Competing on Analytics*, by Tom Davenport
  - “Some companies have built their very businesses on their ability to collect, analyze, and act on data.”
  - “Although numerous organizations are embracing analytics, *only a handful have achieved this level of proficiency. But analytics competitors are the leaders in their varied fields—consumer products finance, retail, and travel and entertainment among them.*”
  - “Organizations are moving beyond query and reporting” - IDC 2006
- *Super Crunchers*, by Ian Ayres
  - “In the past, one could get by on intuition and experience. Times have changed. *Today, the name of the game is data.*”  
—Steven D. Levitt, author of *Freakonomics*
  - “*Data-mining and statistical analysis have suddenly become cool.... Dissecting marketing, politics, and even sports, stuff th complex and important shouldn't be this much fun to read.*” —Wired



# Competitive Advantage



A man in a dark suit, light blue shirt, and striped tie is sitting in an office chair, gesturing with his right hand. He is positioned in front of a large server rack. The server rack has a perforated metal front and various control buttons and indicators on the right side. The background is a blurred office setting with large windows.

# In-Database Data Mining

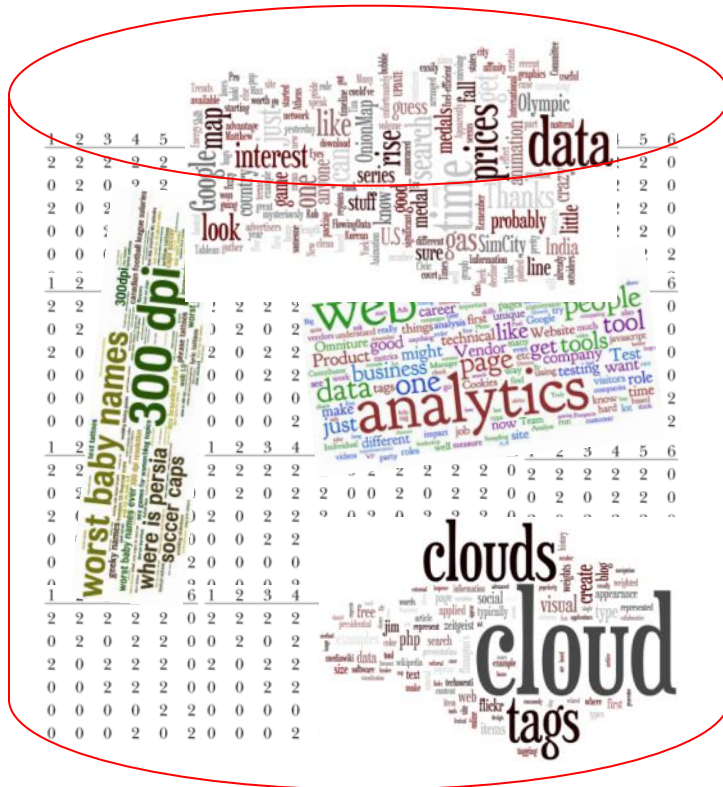
SOFTWARE.  
HARDWARE.  
**COMPLETE.**

ORACLE®

# What is In-Database Analytics?

*Move the data??*

*Move the algorithms?*



$$Z = R * F_P * F_N * F_T * F_C * F_L * F_B$$

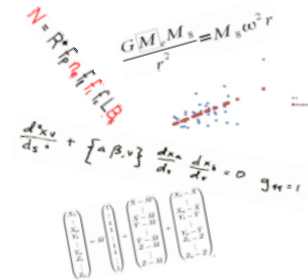
$$\frac{GM_e M_s}{r^2} = M_s \omega^2 r$$

$$\frac{d^2 x_v}{ds^2} + \{a, \beta, v\} \frac{dx_a}{ds} \frac{dx_b}{ds} = 0 \quad g_{rr} = 1$$

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Y_1 \\ \vdots \\ Y_n \\ Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \bar{M} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \bar{X} - \bar{M} \\ \vdots \\ \bar{X} - \bar{M} \\ \bar{Y} - \bar{M} \\ \vdots \\ \bar{Y} - \bar{M} \\ \bar{Z} - \bar{M} \\ \vdots \\ \bar{Z} - \bar{M} \end{pmatrix} + \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \\ Y_1 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \\ Z_1 - \bar{Z} \\ \vdots \\ Z_n - \bar{Z} \end{pmatrix}$$

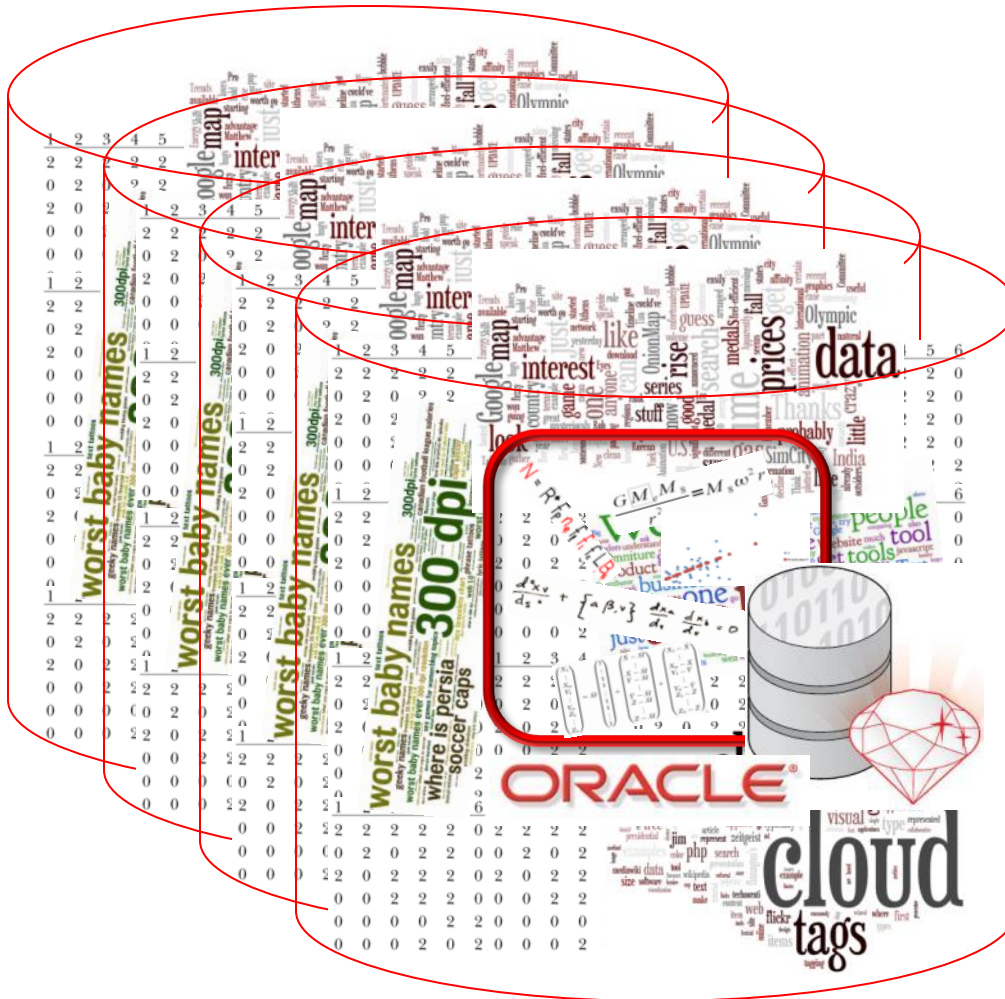


## *Move the algorithms?*



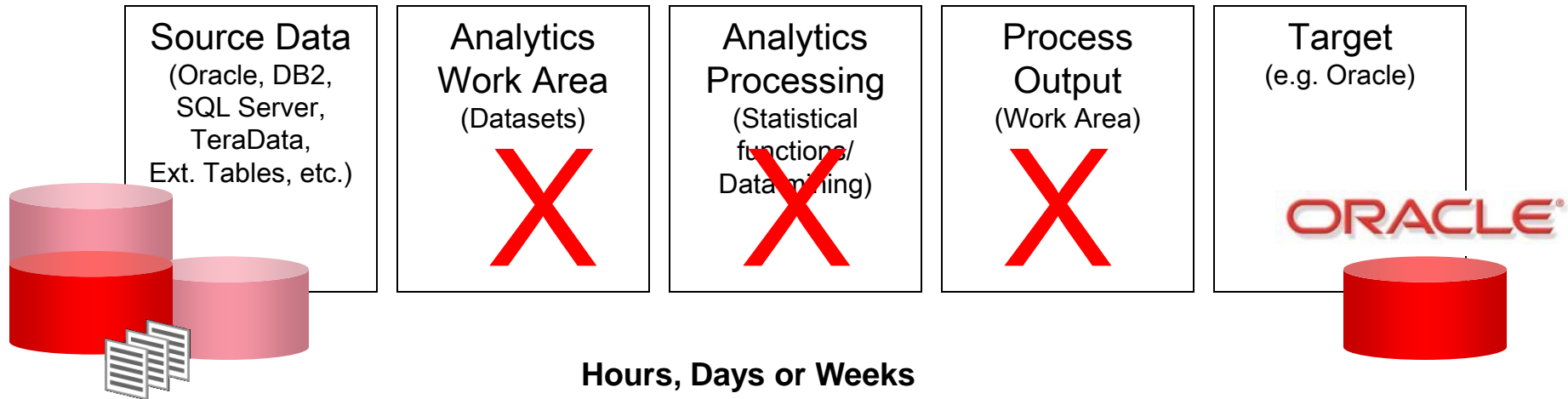
# What is In-Database Analytics?

*Move the algorithms!!!!*



# Traditional Analytics Environment

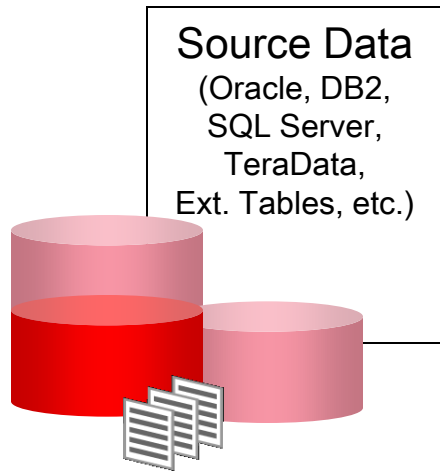
*Move Data → Algorithms*



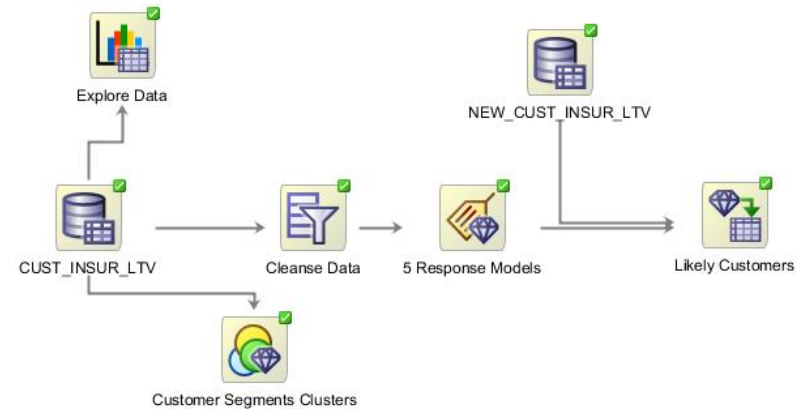
- Traditional analytics environment results in:
  - Data movement
  - Data duplication
  - Loss of security

# Oracle Architecture

## *Move Data ← Algorithms*



**Secs, Mins or Hours**



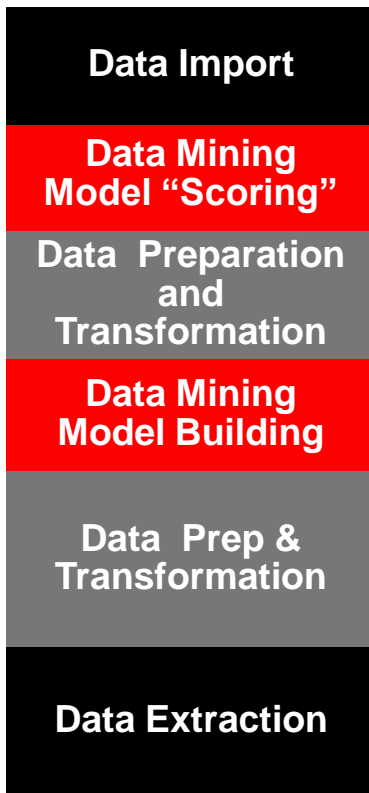
- Oracle architecture:
  - Eliminates data movement
  - Eliminates data duplication
  - Preserves security



# In-Database Data Mining



## Traditional Analytics



## Oracle Data Mining

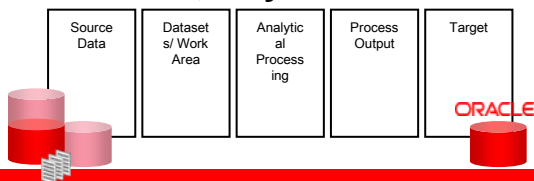
**\$avings**

### Results

- Faster time for “Data” to “Insights”
- Lower TCO—Eliminates
  - Data Movement
  - Data Duplication
- Maintains Security

- Model “Scoring”  
Data remains in the Database
- Embedded data preparation
- Cutting edge machine learning algorithms inside the SQL kernel of Database
- SQL—Most powerful language for data preparation and transformation
- Data remains in the Database

Hours, Days or Weeks



Secs. Mins or Hours



ORACLE

A man in a dark suit, light blue shirt, and striped tie is sitting in an office chair, gesturing with his right hand. He is positioned in front of a large server rack. The server rack has a perforated metal front and various control buttons and indicators on the right side. The background is a blurred office setting with large windows.

# Oracle Data Mining

SOFTWARE.  
HARDWARE.  
**COMPLETE.**

ORACLE®



- 11 years “stem celling analytics” into Oracle
  - Designed advanced analytics into database kernel to leverage relational database strengths
  - Naïve Bayes and Association Rules—1<sup>st</sup> algorithms added
  - Leverages counting, conditional probabilities, and much more
- Now, analytical database platform
  - 12 cutting edge machine learning algorithms and 50+ statistical functions
  - A data mining model is a schema object in the database, built via a PL/SQL API and scored via built-in SQL functions.
  - When building models, leverage existing scalable technology
    - (e.g., parallel execution, bitmap indexes, aggregation techniques) and add new core database technology (e.g., recursion within the parallel infrastructure, IEEE float, etc.)
  - True power of embedding within the database is evident when scoring models using built-in SQL functions (incl. Exadata)

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

# You Can Think of It Like This...

## Traditional SQL

- “Human-driven” queries
- Domain expertise
- Any “*rules*” must be defined and managed
- SQL Queries
  - SELECT
  - DISTINCT
  - AGGREGATE
  - WHERE
  - AND OR
  - GROUP BY
  - ORDER BY
  - RANK



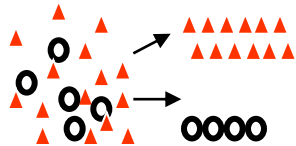
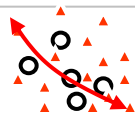
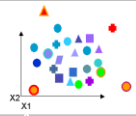
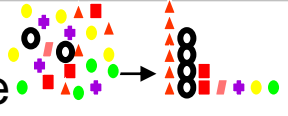
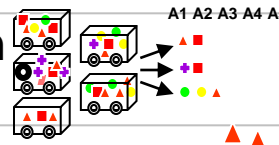
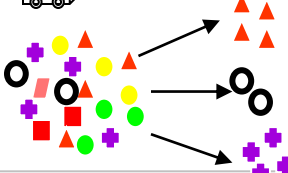
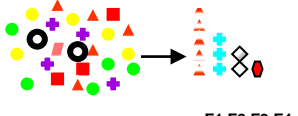
## Oracle Data Mining

- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “*right*” data to mine
- ODM “Verbs”
  - PREDICT
  - DETECT
  - CLUSTER
  - CLASSIFY
  - REGRESS
  - PROFILE
  - IDENTIFY FACTORS
  - ASSOCIATE





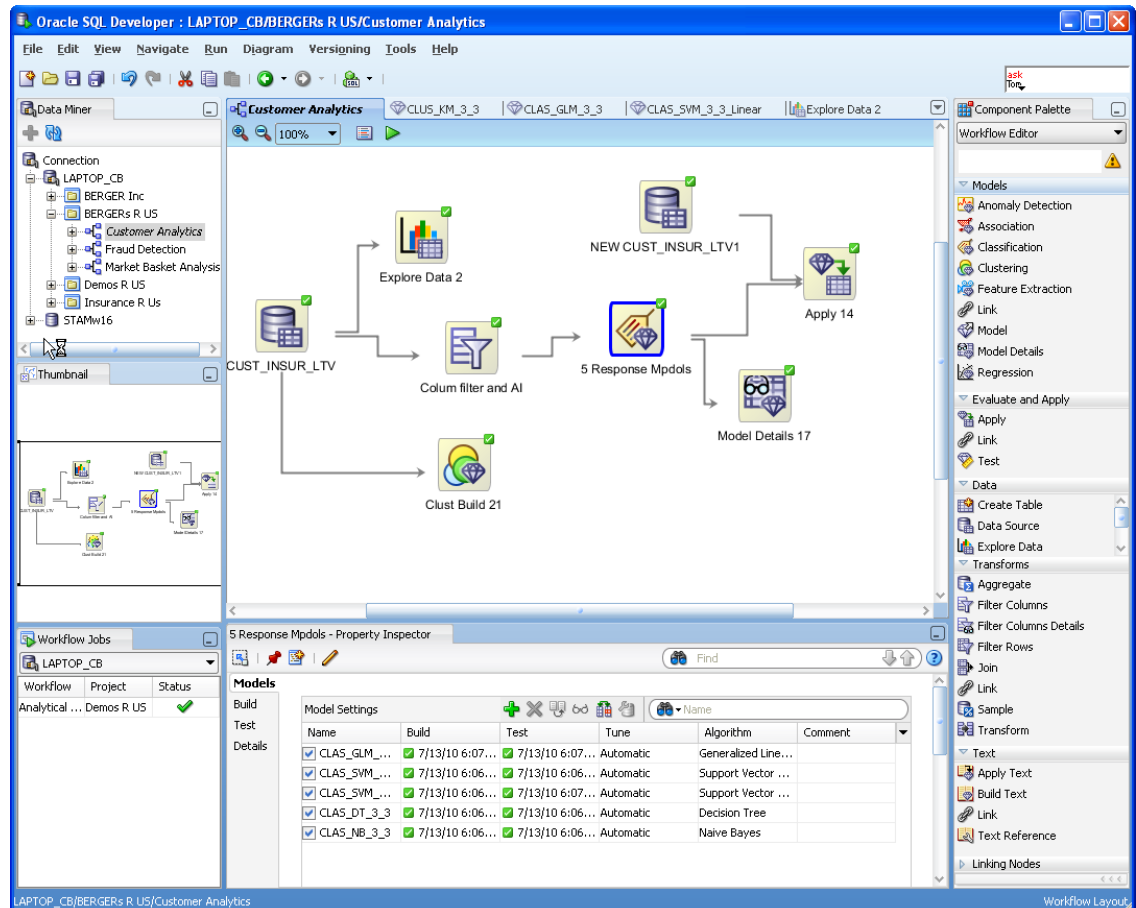
# Oracle Data Mining Algorithms

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 	NMF	Text analysis Feature reduction

# Oracle Data Miner 11g Release 2 GUI

## *Free SQL Developer Extension on OTN*

- 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 Data Miner 11g Release 2 GUI

## Free SQL Developer Extension on OTN

The image displays four screenshots of the Oracle Data Miner 11g Release 2 GUI, illustrating the workflow for targeting best customers.

**Top Left Screenshot:** Shows the 'Targeting Best Customers' workflow. The process flow is: Explore Data (CUST\_INSUR\_LTV) → Cleanse Data → 5 Response Models → Likely Customers (NEW\_CUST\_INSUR\_LTV).

**Top Right Screenshot:** Shows the 'Response Models' window for 'CLAS\_DT\_5\_1'. It displays a decision tree with nodes 8, 9, 6, 19, 20, 21, 22, 17, and 18. Each node shows prediction statistics (Yes/No counts and percentages) and confidence levels.

**Bottom Left Screenshot:** Shows the 'Explore Data 2' window for 'CUST\_INSUR\_LTV'. It displays a table of statistics grouped by 'LTV\_BIN'.

Name	Histogram	Data Type	Percent NULLs	Distinct Values	Mode
"CREDIT_CARD_LIMITS"		NUMBER	0	28	
"CHECKING_AMOUNT"		NUMBER	0	636	
"MARITAL_STATUS"		VARCHAR2	0	5	MARRIED
"LTV"		NUMBER	0	1,930	
"HAS_CHILDREN"		NUMBER	0	2	
"LAST"		VARCHAR2	0	1,396	EMERY
"MONEY_MONTHLY_OVERDRAWN"		NUMBER	0	371	
"STATE"		VARCHAR2	0	24	NY
"SALARY"		NUMBER	0	1,906	
"BANK_FUNDS"		NUMBER	0	446	
"MORTGAGE_AMOUNT"		NUMBER	0	413	

Below the table is a bar chart titled '"MARITAL\_STATUS By LTV\_BIN"'. The x-axis shows marital status categories: DIVORCED, MARRIED, OTHER, SINGLE, and WIDOWED. The y-axis shows Percent (0 to 100). The legend indicates three LTV categories: LOW (blue), 'VERY HIGH' (green), and 'MEDIUM' (red).

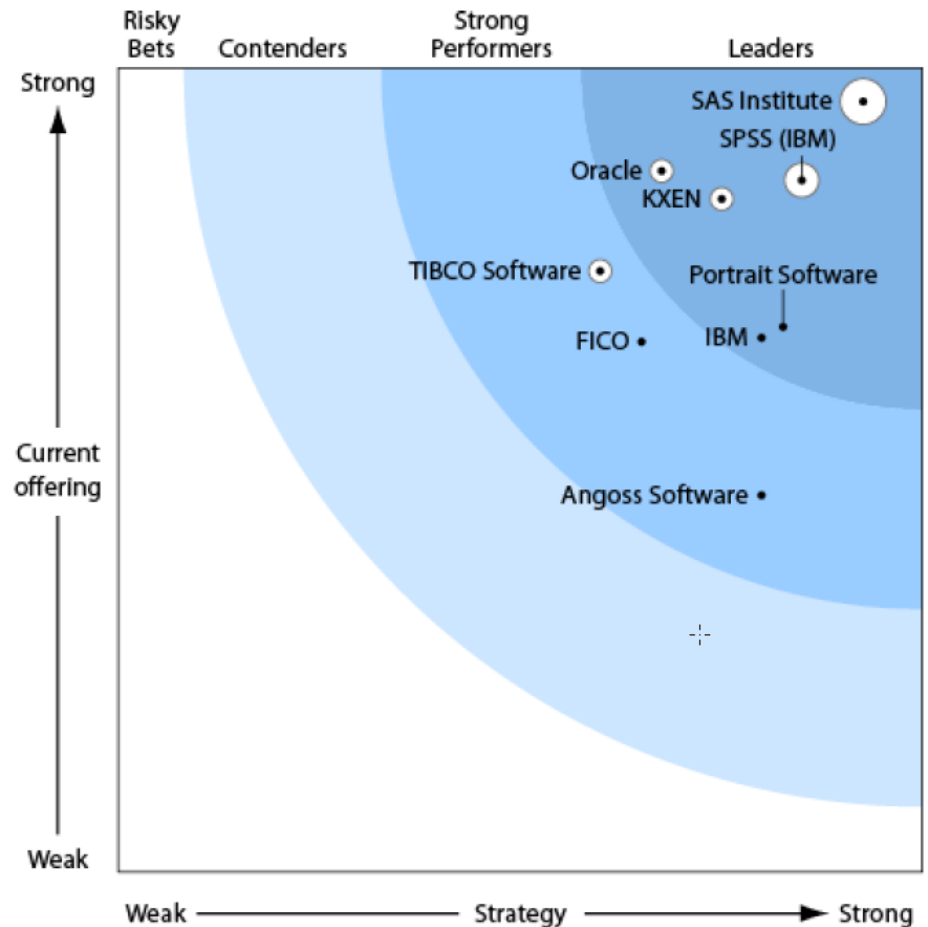
**Bottom Right Screenshot:** Shows the 'Target Values' window for 'dataminer.svmclassification'. It displays a list of features and their corresponding target values.

Feature	Target Value
AGE	66.6463722
N_OF_DEPENDENTS	0.86557898
MORTGAGE_AMOUNT	-0.85111609
STATE	-0.76673662
STATE	0.68542101
SALARY	-0.55453918
STATE	-0.48900215
HOUSE_OWNERSHIP	-0.34647433
STATE	-0.34306028
N_TRANS_KIOSK	0.32794196
STATE	0.32315484
STATE	0.30801241
HOUSE_OWNERSHIP	0.28906774
TIME_AS_CUSTOMER	-0.27872530
TIME_AS_CUSTOMER	0.24943188

# The Forrester Wave™: Predictive Analytics And Data Mining Solutions, Q1 2010

*Oracle Data Mining Cited as a Leader; 2<sup>nd</sup> place in Current Offering*

- Ranks 2<sup>nd</sup> place in Current Offering
- “Oracle focuses on in-database mining in the Oracle Database, on integration of Oracle Data Mining into the kernel of that database, and on leveraging that technology in Oracle’s branded applications.”



The Forrester Wave is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave are trademarks of Forrester Research, Inc. The Forrester Wave is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings, and comments. Forrester does not endorse any vendor, product, or service depicted in the Forrester Wave. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.



A man in a dark suit, light blue shirt, and striped tie is sitting in a black leather office chair. He is gesturing with his right hand, palm facing up. Behind him is a large Oracle Exadata database machine, which is a tall, silver-colored server rack with a perforated front panel. The machine has various labels and buttons, including "TAPE", "STANDBY", and "XSCF".

# Exadata & ODM

SOFTWARE.  
HARDWARE.  
**COMPLETE.**

ORACLE®

# Exadata + Data Mining 11g Release 2

## *“DM Scoring” Pushed to Storage!*



- In 11g Release 2, SQL predicates and Oracle Data Mining models are pushed to storage level for execution

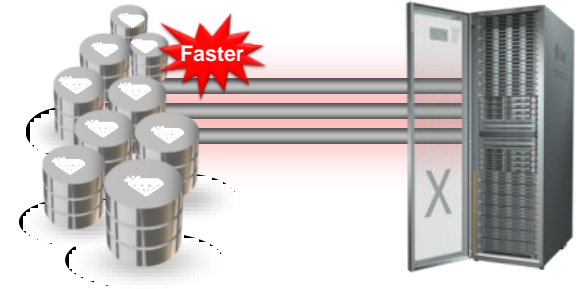
For example, find the US customers likely to churn:

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

Scoring function executed in Exadata

# Exadata + Data Mining 11g Release 2

## Benefits



- Eliminates data movement
  - 2X-5X+ faster scoring on Exadata
    - Depends on number of joins involved with data for scoring
- Preserves security
- Significant architecture and performance advantages over SAS Institute
  - Years ahead of SAS's road map to move SAS analytics towards RDBMSs (<http://support.sas.com/resources/papers/InDatabase07.pdf>)
- Netezza performance but using industry standard RDBMS + SQL-based in-database advanced analytics
- Best platform for building enterprise predictive analytics applications e.g. Fusion Applications →  
“Analytical iPod for the Enterprise”

# TurkCell Prepaid Churn Model

## Oracle Data Mining on Exadata 11g Release 2



- Churn Problem
  - Churn prediction starts with turning an abundance of data into valuable information and continues as a cyclic process
- Approach
  - Initially we have used a large Solaris (100+ UltraSparc 7 cores and 640 GB memory) box to build our first SVM models:
  - It took 29 hours to complete model build & apply.
- Conclusion
  - On Exadata this reduces to a few hours mainly due to enormous improvement in data preparation stage
  - Churn prediction over various customer groups is and will be the focus of Turkcell
  - Embedded data mining with ODM is faster, more robust (due to stability of SVM algorithm), easier to automate, easier to manage

Excerpts from TurkCell presentation at OOW 2010, September 21, 2010  
Necdet Deniz Halicioğlu [deniz.halicioğlu@turkcellteknoloji.com.tr](mailto:deniz.halicioğlu@turkcellteknoloji.com.tr)







Easier

# Oracle Data Miner 11g Release 2

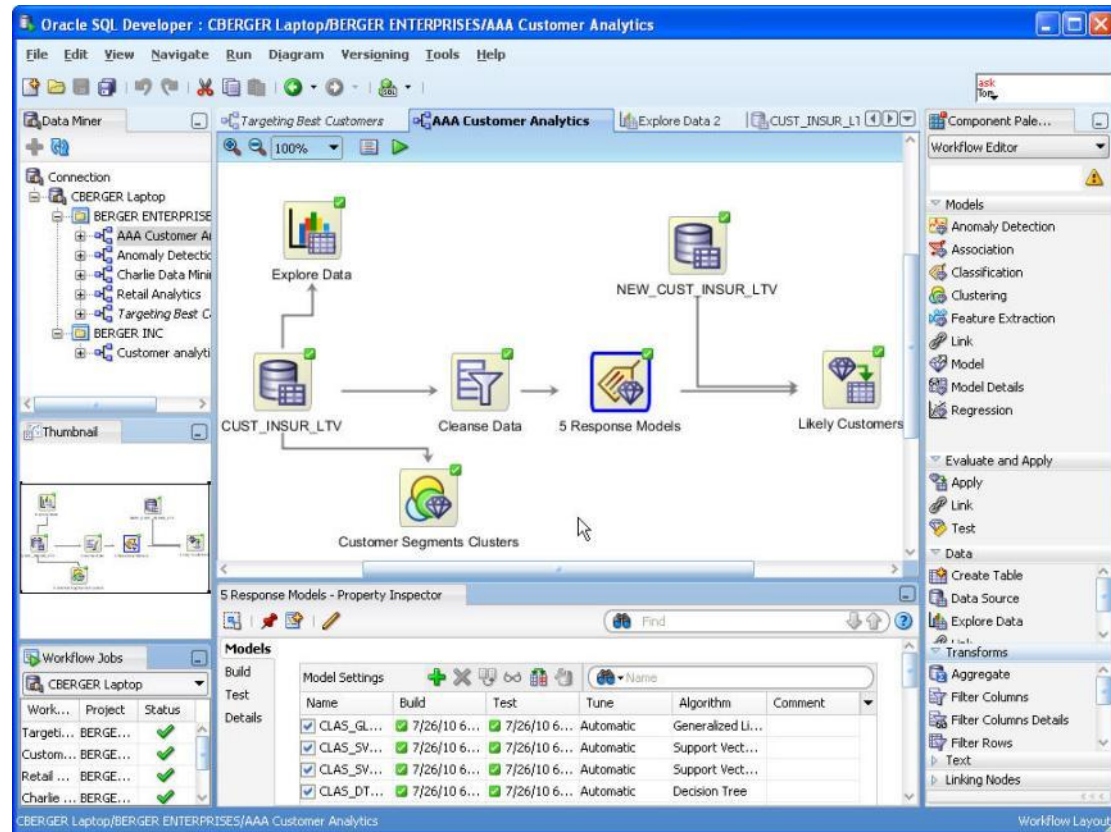
SOFTWARE.  
HARDWARE.  
COMPLETE.



ORACLE

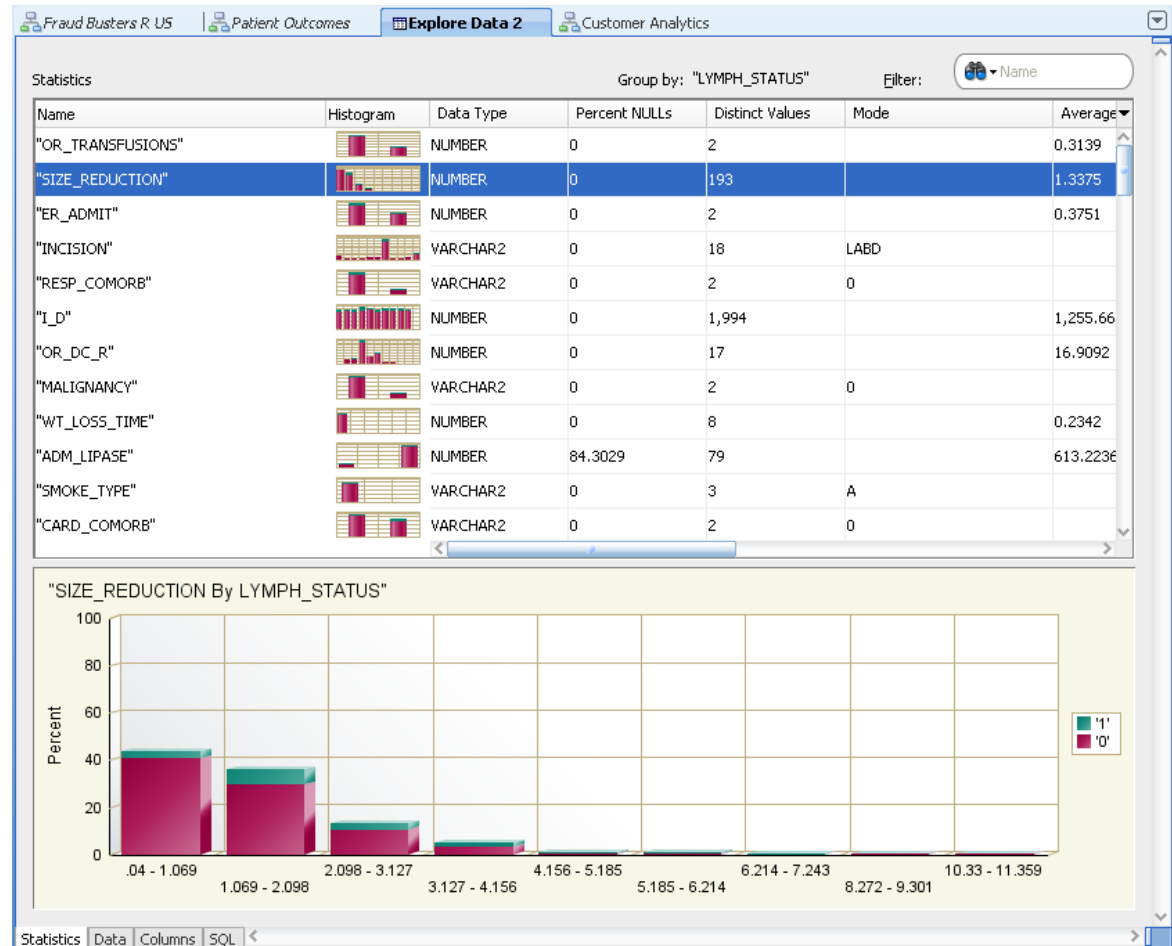
# Oracle Data Miner 11g Release 2 GUI

- Predict customer behavior
- Identify key factors
- Predict next-likely product
- Customer profiling
- Detect fraud & anomalies
- Mine “text” and unstructured data



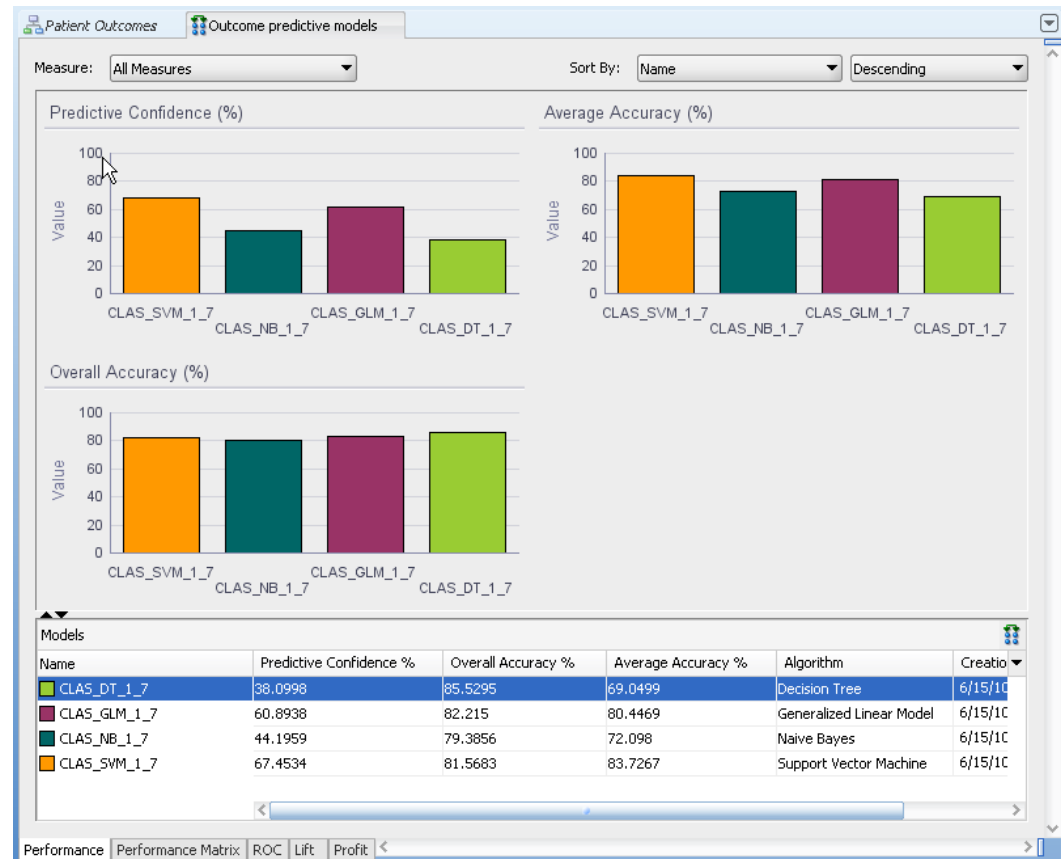
# Explore Data

- Thumbnail distributions of every attribute
  - Grouped by another attribute
- Summary statistics for all attributes
  - Min, max, stdev, variance, median, mean, skewness, kurtosis, etc.



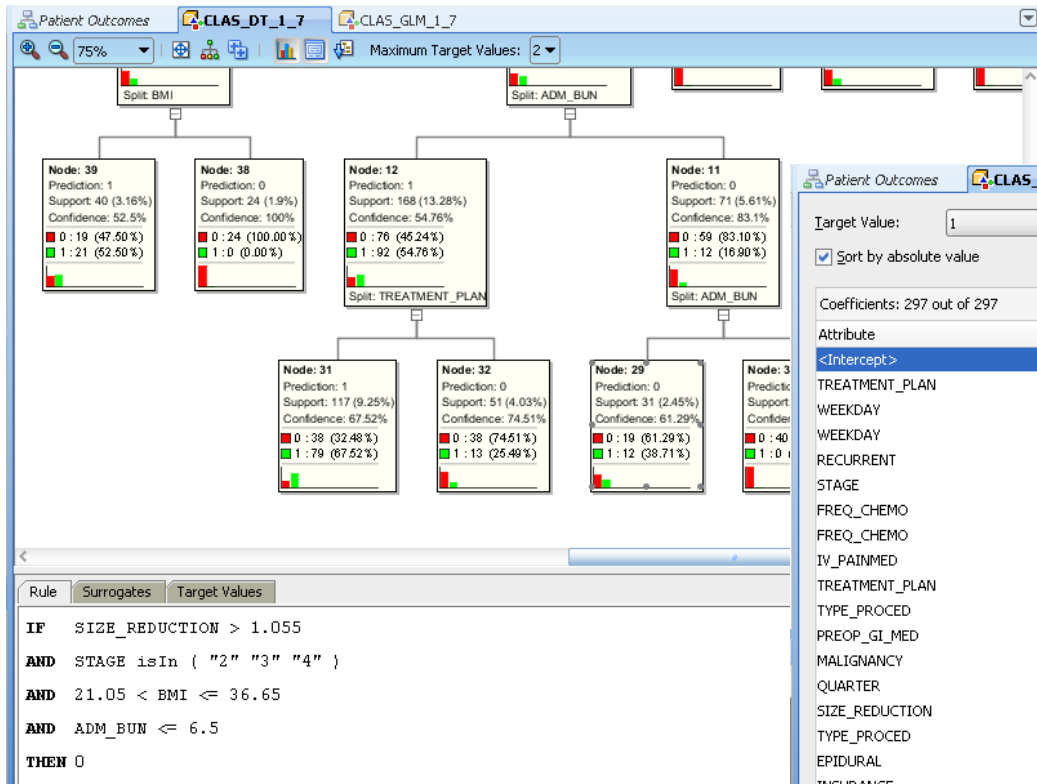
# Build and Evaluate Models

- Comparative model performance results
- Adjust and tune predictive models



# Understand Model Details

- Interactive model viewers



Outcome predictive models

Target Value: 1

Sort by absolute value

Fetch Size: 10,000

Coefficients: 297 out of 297

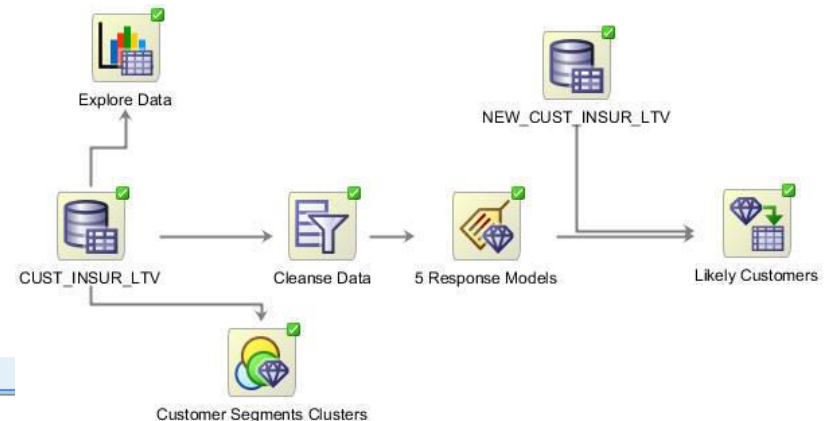
Attribute	Value	Coefficient	Standardized Coeffi...	Exp(Coefficient)
<Intercept>	NULL	-1.83481346	0	6.26396556
TREATMENT_PLAN	Chemo_only	-0.46513283	0.11735002	1.59222567
WEEKDAY	W	-0.40697858	0.0869471	1.50227193
WEEKDAY	Th	-0.34941526	0.05883753	1.418238
RECURRENT	1	-0.33993936	0.07348783	1.4048624
STAGE	3	0.29916993	-0.06150948	0.74143341
FREQ_CHEMO	1	0.29378459	-0.06262496	0.74543705
FREQ_CHEMO	0	-0.26376819	0.05597178	1.30182638
IV_PAINMED	DEM	-0.26085980	0.036163	1.29804567
TREATMENT_PLAN	Chemo&Radiation	-0.25534174	0.03324906	1.2909027
TYPE_PROCD	closed	0.25466832	-0.01992872	0.77517356
PREOP_GI_MED	1	0.25194913	-0.06873117	0.77728428
MALIGNANCY	1	0.24061736	-0.05486614	0.78614238
QUARTER	A	0.23306129	-0.05746447	0.79210502
SIZE_REDUCTION	NULL	0.22915110	-0.15356344	0.79520837
TYPE_PROCD	1	-0.22759025	0.03846051	1.25557075
EPIDURAL	1	-0.22715954	0.05119796	1.25503009
INSURANCE	B	0.21168257	-0.05517357	0.80922152
OR_TRANSFUSIONS	1	0.20613024	-0.0550411	0.81372709
TYPE_ABX	Cipro	0.20248206	-0.02044382	0.81670114
EKG	SB	0.19228831	-0.02216336	0.82506896
IV_PAINMED	TORD	-0.19105185	0.01912802	1.21052222
INCISION	KNEE	-0.18882816	0.01878139	1.20783338
INSURANCE	C	0.18859100	-0.02710814	0.82812514
WT_LOSS_TIME	NULL	-0.17535293	0.11368976	1.19166672
WEEKDAY	Sa	0.17096336	-0.02674837	0.84285246

Details Coefficients Compare Settings



# Analytical “Work Flow” Methodologies

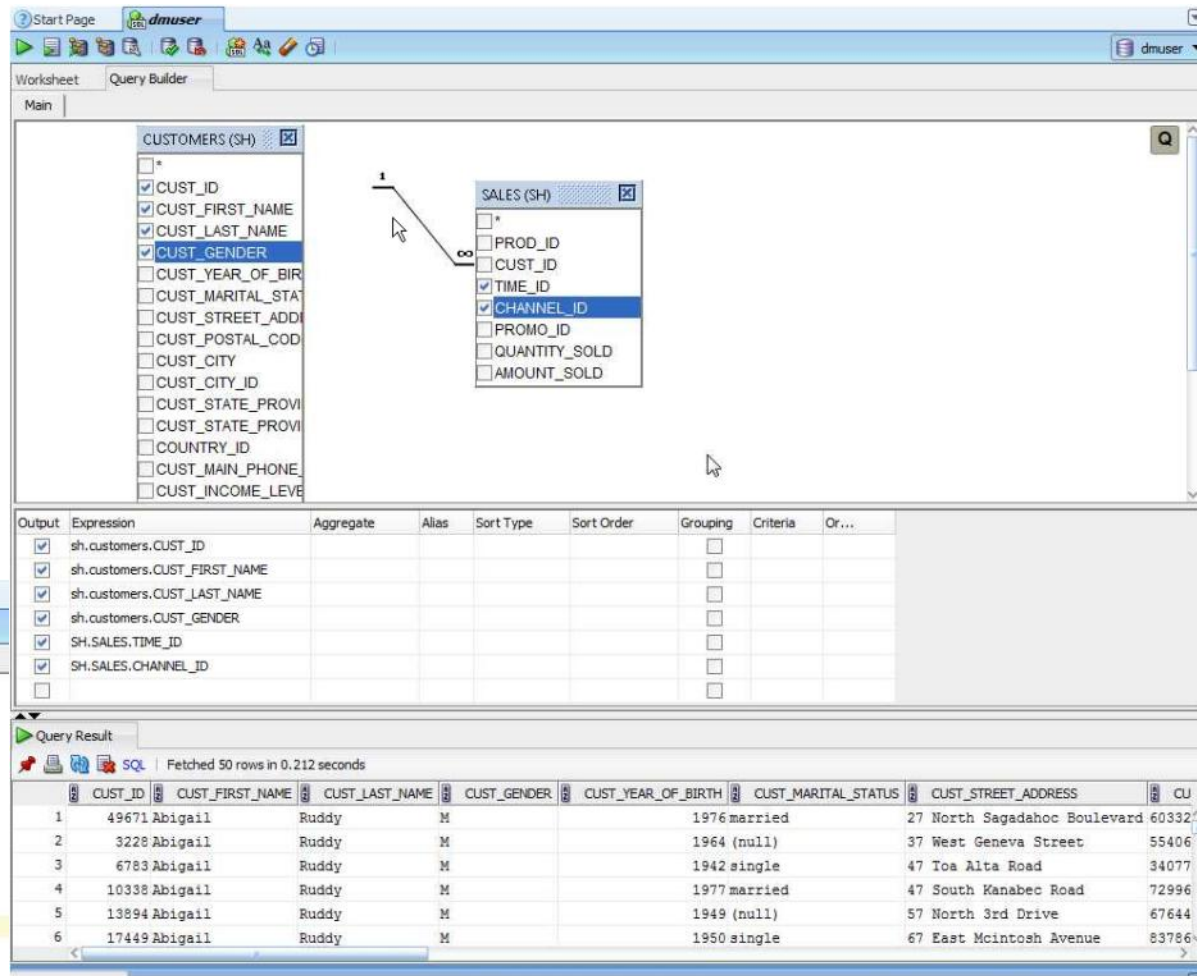
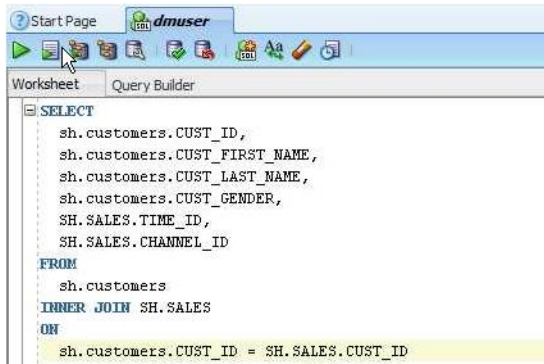
- *Build, share and automate predictive analytics methodologies*



Patient Outcomes									
At Risk patients									
CLAS_DT_1_7									
CLAS_GLM_1_7									
View: Cache Data   Sort...   Filter: Enter Where Clause									
Customer Segments Clusters									
	CLAS_SVM_1_7_PRED	CLAS_SVM_1_7_PROB	LYMPH_TYPE	SIZE_TUMOR_MM	MARITAL	ADM_ALBUMIN	AMT_CHEMO	FREQ_CHEMO	CLAS_DT_1_7
1	1	0.99865991	Agressive	7,100	M	1.6	42.17	1	1
2	1	0.99865991	Agressive	7,100	M	1.6	42.17	1	1
3	1	0.99865991	Agressive	7,100	M	1.6	42.17	1	1
4	1	0.99351446	Agressive	5,200	M	2.4	52	1	1
5	1	0.99351446	Agressive	5,200	M	2.4	52	1	1
6	1	0.99351446	Agressive	5,200	M	2.4	52	1	1
7	1	0.99149541	Agressive	1,350	S	2.4	37.01	2	2
8	1	0.99149541	Agressive	1,350	S	2.4	37.01	2	2
9	1	0.99149541	Agressive	1,350	S	2.4	37.01	2	2
10	1	0.99149541	Agressive	1,350	S	2.4	37.01	2	2
11	1	0.9912111	Indolent	3,400	W		3.25	2	2
12	1	0.9912111	Indolent	3,400	W		3.25	2	2
13	1	0.9912111	Indolent	3,400	W		3.25	2	2
14	1	0.9912111	Indolent	3,400	W		3.25	2	2
15	1	0.98217842	Indolent	1,000	M	2.4	52.25	1	1
16	1	0.98217842	Indolent	1,000	M	2.4	52.25	1	1

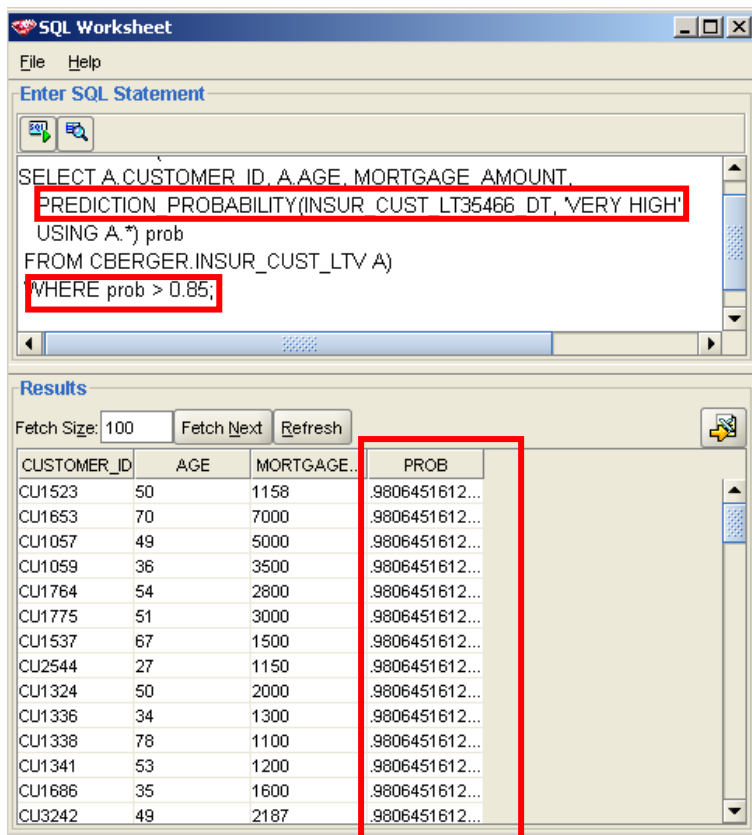
# SQL Developer Active Query Builder

- New, easy to use, interactive query builder in SQL Developer for assembling and preparing data—for mining*



# Example: Simple, Predictive SQL

Select customers who are **more than 85% likely to be HIGH VALUE customers** & display their AGE & MORTGAGE\_AMOUNT



The screenshot shows an SQL Worksheet window with a query entered in the 'Enter SQL Statement' pane. The query is: `SELECT A.CUSTOMER_ID, A.AGE, MORTGAGE_AMOUNT, PREDICTION_PROBABILITY(INSUR_CUST_LT35466_DT, 'VERY HIGH' USING A.*) prob FROM CBERGER.INSUR_CUST_LTV A) WHERE prob > 0.85;`. The 'Results' pane below shows a table with four columns: CUSTOMER\_ID, AGE, MORTGAGE\_AMOUNT, and PROB. The table contains 15 rows of data. The 'PROB' column values are all truncated to '.9806451612...'. A red box highlights the 'PROB' column in the results table.

CUSTOMER_ID	AGE	MORTGAGE_AMOUNT	PROB
CU1523	50	1158	.9806451612...
CU1653	70	7000	.9806451612...
CU1057	49	5000	.9806451612...
CU1059	36	3500	.9806451612...
CU1764	54	2800	.9806451612...
CU1775	51	3000	.9806451612...
CU1537	67	1500	.9806451612...
CU2544	27	1150	.9806451612...
CU1324	50	2000	.9806451612...
CU1336	34	1300	.9806451612...
CU1338	78	1100	.9806451612...
CU1341	53	1200	.9806451612...
CU1686	35	1600	.9806451612...
CU3242	49	2187	.9806451612...

```
SELECT * from(  
SELECT A.CUST_ID, A.AGE,  
MORTGAGE_AMOUNT, PREDICTION_PROBABILITY  
(CUST_INSUR_LT46939_DT, 'VERY HIGH'  
USING A.*) prob  
FROM CBERGER.CUST_INSUR_LTV A)  
WHERE prob > 0.85;
```

# 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',  
    'CLAIMS2', '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 CLAIMS2  
where PASTNUMBEROFCLAIMS in ('2 to 4', 'more than 4'))  
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



# Real-time Prediction

with

```
records as (select
  78000  SALARY,
  250000 MORTGAGE_AMOUNT,
  6  TIME_AS_CUSTOMER,
  12  MONTHLY_CHECKS_WRITTEN,
  55  AGE,
  423  BANK_FUNDS,
  'Married'  MARITAL_STATUS,
  'Nurse'  PROFESSION,
  'M'  SEX,
  4000  CREDIT_CARD_LIMITS,
  2  N_OF_DEPENDENTS,
  1  HOUSE_OWNERSHIP from dual)
```

```
select s.prediction prediction, s.probability probability
```

```
from (
```

```
  select PREDICTION_SET(CUST_INSUR_LT46939_DT, 1 USING *) pset
  from records) t, TABLE(t.pset) s;
```

**On-the-fly, single record  
apply with new data (e.g.  
from call center)**

PREDICTION	PROBABILITY
HIGH	.65123504738232096



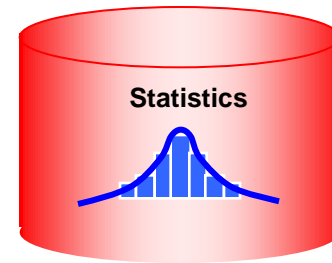
A man in a dark suit, light blue shirt, and striped tie is sitting in an office chair, gesturing with his right hand. He is positioned in front of a row of server racks. The server racks have perforated metal doors and various control panels with buttons and indicators. The background is a blurred office setting with large windows.

# Oracle Statistical Functions (Free)

SOFTWARE.  
HARDWARE.  
**COMPLETE.**

ORACLE®

# 11g Statistics & SQL Analytics (Free)

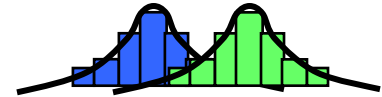


- Ranking functions
  - rank, dense\_rank, cume\_dist, percent\_rank, ntile
- Window Aggregate functions (moving and cumulative)
  - Avg, sum, min, max, count, variance, stddev, first\_value, last\_value
- LAG/LEAD functions
  - Direct inter-row reference using offsets
- Reporting Aggregate functions
  - Sum, avg, min, max, variance, stddev, count, ratio\_to\_report
- Statistical Aggregates
  - Correlation, linear regression family, covariance
- Linear regression
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR\_POP, COVAR\_SAMP, and CORR functions

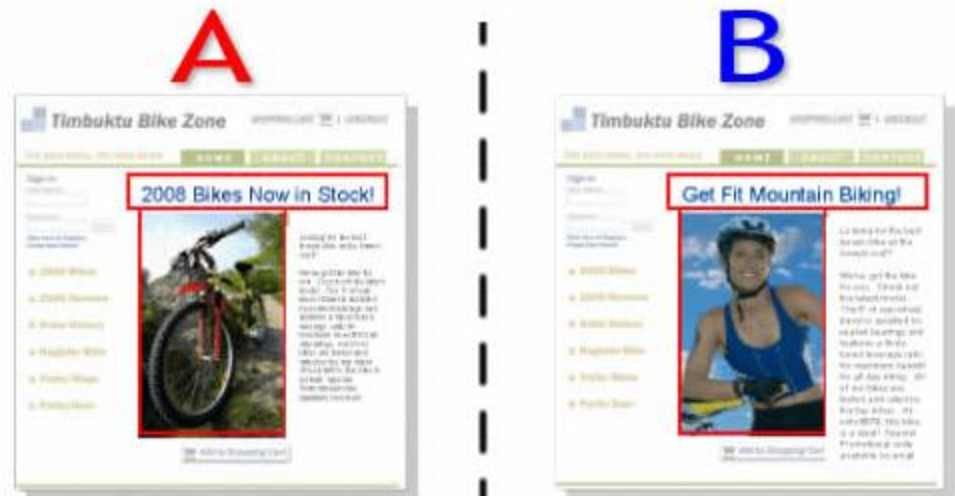
## Descriptive Statistics

- DBMS\_STAT\_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, median, stats\_mode, variance, standard deviation, quantile values, +/- n sigma values, top/bottom 5 values
- Correlations
  - Pearson's correlation coefficients, Spearman's and Kendall's (both nonparametric).
- Cross Tabs
  - Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa
- Hypothesis Testing
  - Student t-test, F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA
- Distribution Fitting
  - Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential

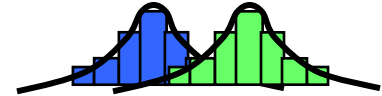
# Split Lot A/B Offer testing



- Offer “**A**” to one population and “**B**” to another
- Over time period “**t**” calculate **median** purchase amounts of customers receiving offer **A** & **B**
- Perform **t-test** to compare
- If statistically significantly better results achieved from one offer over another, offer everyone higher performing offer



# Independent Samples T-Test (Pooled Variances)



- Query compares the mean of AMOUNT\_SOLD between MEN and WOMEN within CUST\_INCOME\_LEVEL ranges

```
SELECT substr(cust_income_level,1,22) income_level,  
       avg(decode(cust_gender, 'M', amount_sold, null)) sold_to_men,  
       avg(decode(cust_gender, 'F', amount_sold, null)) sold_to_women,  
       stats_t_test_indep(cust_gender, amount_sold, 'STATISTIC', 'F')  
       t_observed,  
       stats_t_test_indep(cust_gender, amount_sold) two_sided_p_value  
FROM sh.customers c, sh.sales s  
WHERE c.cust_id=s.cust_id  
GROUP BY rollup(cust_income_level)  
ORDER BY 1;
```

**SQL Worksheet**





**Ability to Import  
3<sup>rd</sup> Party e.g. SAS Models**

**New**

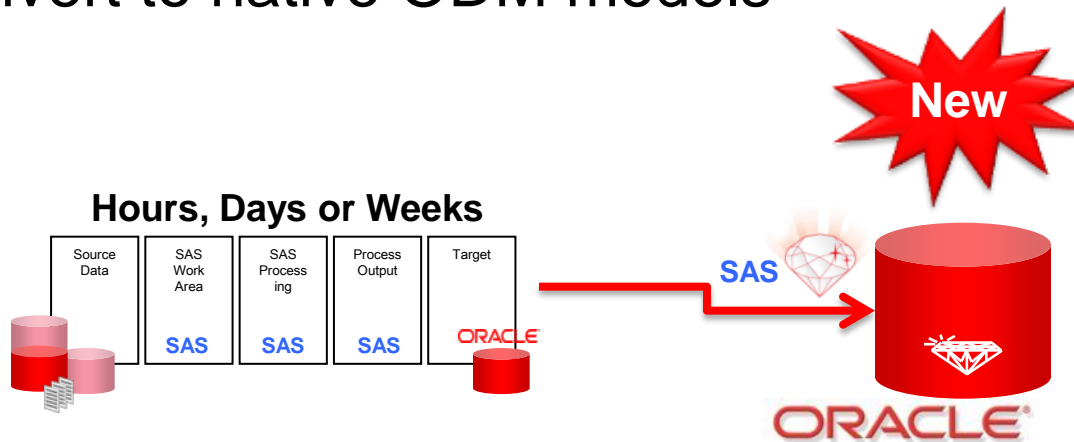
**SOFTWARE.  
HARDWARE.  
COMPLETE.**

**ORACLE®**



# Ability to Import 3<sup>rd</sup> Party DM Models

- Capability to import 3<sup>rd</sup> party dm models, import, and convert to native ODM models

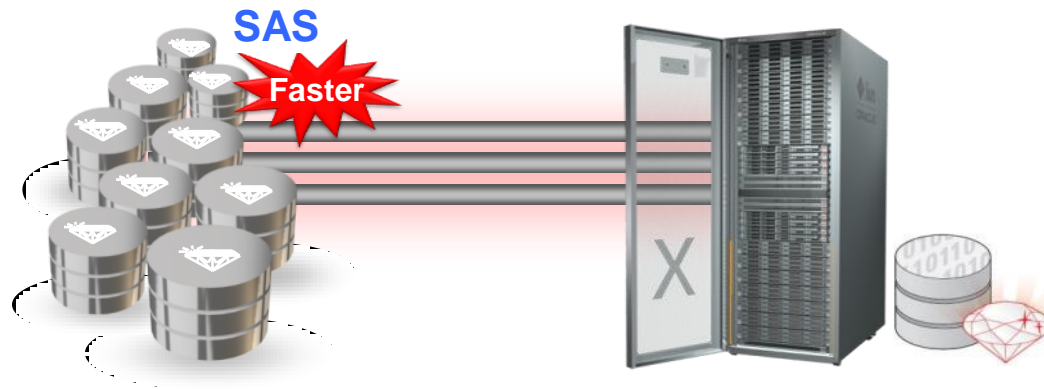
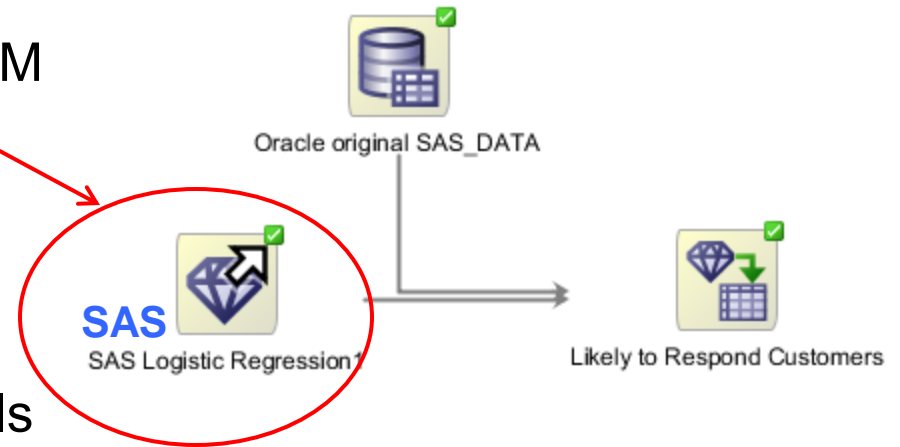


- Benefits
  - SAS, SPSS, R, etc. data mining models can be used for scoring inside the Database
  - Imported dm models become native ODM models and inherit all ODM benefits including scoring at Exadata storage layer, 1<sup>st</sup> class objects, security, etc.

# In-Database SAS Scoring

## *Score the SAS\_ODM Model*

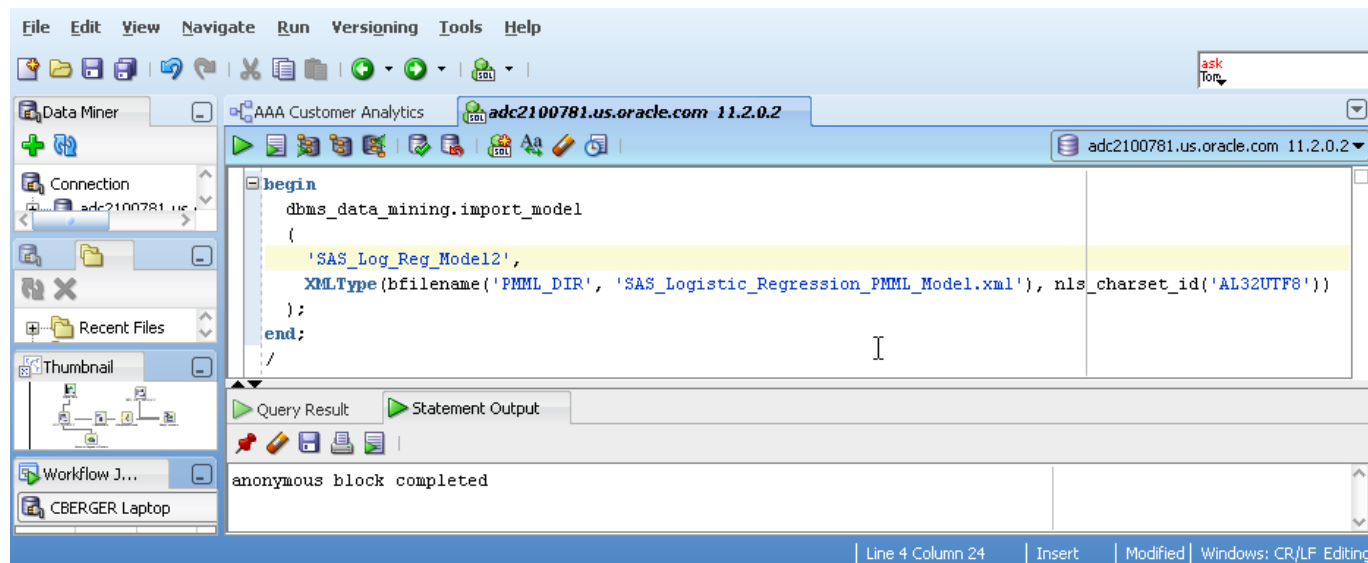
- SAS models become native ODM models
  - No loss of information
- Original source data for scoring remains in Database
- “Exadata scoring” of SAS models



# In-Database SAS Scoring

## *Import the SAS Model*

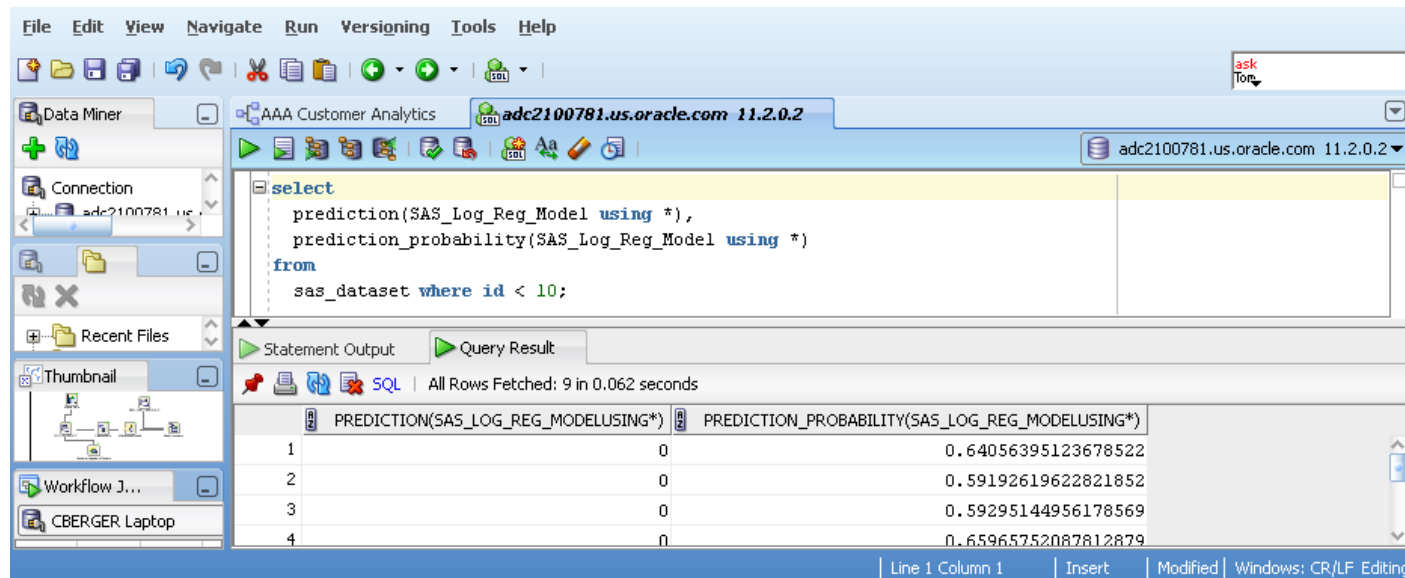
```
begin
  dbms_data_mining.import_model
  (
    'SAS_Log_Reg_Model4',
    XMLType(bfilename('PMML_DIR',
'SAS_Logistic_Regression_PMML_Model.xml'),
nls_charset_id('AL32UTF8'))
  );
end;
/
```



# In-Database SAS Scoring

## *Score the SAS\_ODM Model*

```
select
  prediction(SAS_Log_Reg_Model4 using *),
  prediction_probability(SAS_Log_Reg_Model using *)
from
  sas_dataset where id < 10;
```



The screenshot displays the Oracle SQL Developer environment. The main editor window contains the following SQL query:

```
select
  prediction(SAS_Log_Reg_Model using *),
  prediction_probability(SAS_Log_Reg_Model using *)
from
  sas_dataset where id < 10;
```

The query has been executed, and the results are displayed in the 'Query Result' pane. The results show 4 rows of data, with the first column being the model name and the second column being the prediction probability.

	PREDICTION(SAS_LOG_REG_MODELUSING*)	PREDICTION_PROBABILITY(SAS_LOG_REG_MODELUSING*)
1	0	0.64056395123678522
2	0	0.59192619622821852
3	0	0.59295144956178569
4	0	0.659655752087812879

A man in a dark suit, light blue shirt, and striped tie is sitting in an office chair, gesturing with his right hand. He is positioned in front of a large server rack. The server rack has a perforated metal front and various control buttons and indicators on the right side. A red starburst graphic is placed near the man's hand.

# Applications Powered by Oracle Data Mining

**Simpler!**

**SOFTWARE.  
HARDWARE.  
COMPLETE.**

**ORACLE®**



# Integration with Oracle BI EE

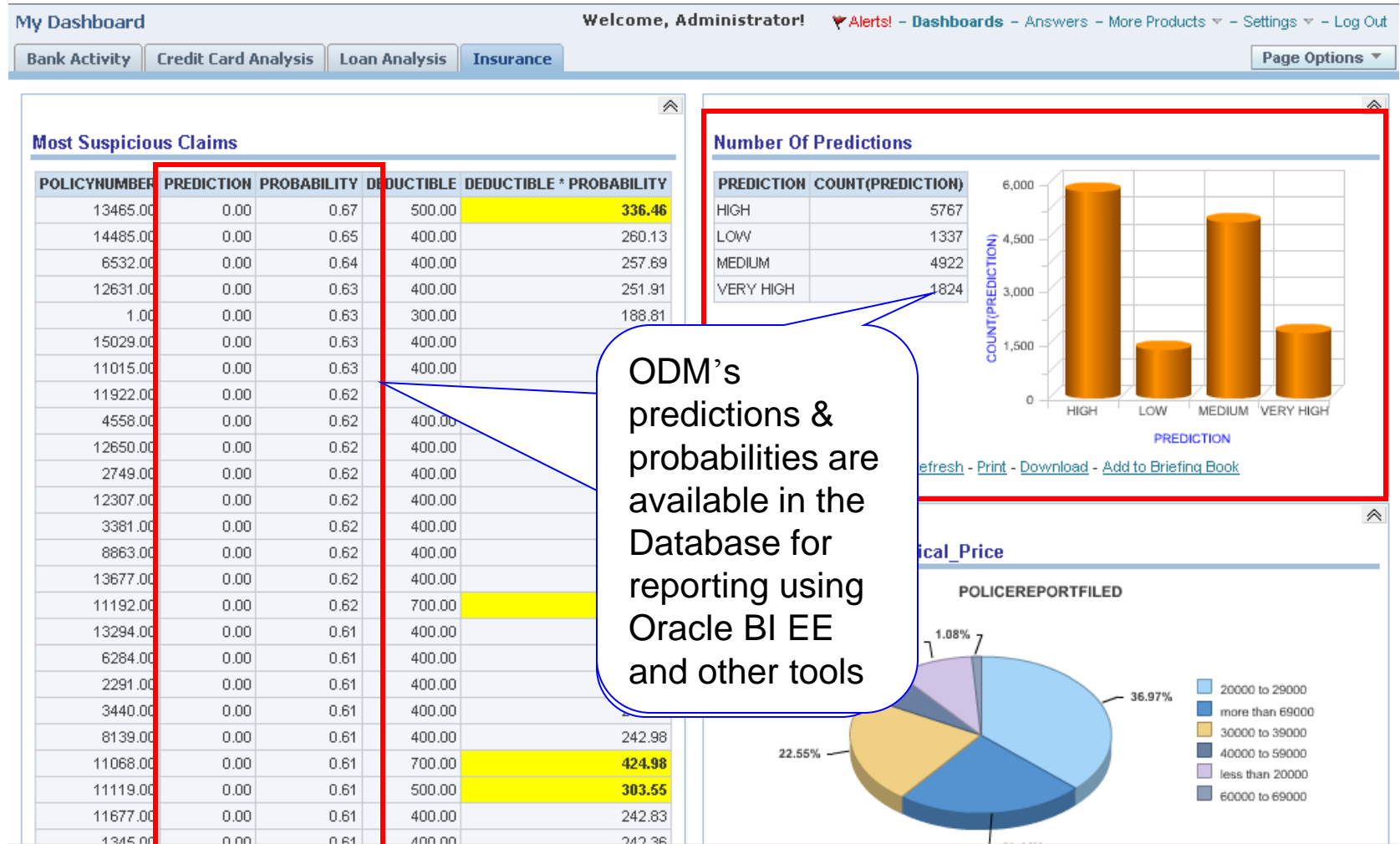
The screenshot displays the (Online) Siebel Analytics Administration Tool - AnalyticsWeb interface, which is divided into three main panels: Presentation, Business Model and Mapping, and Physical.

- Presentation Panel:** This panel shows the hierarchy of data sources and facts. A red circle highlights the **KEY\_FACTOR** and **IMPORTANCE** dimensions under the **CD\_BUYERS** fact. Another red circle highlights the **YRS\_RESIDENCE** dimension under the **FACT** dimension. A callout box points to these dimensions with the text: "Oracle BI EE defines results for end user presentation".
- Business Model and Mapping Panel:** This panel shows the mapping of data sources to the presentation layer. It includes a tree view of the **CD\_BUYERS** fact and its dimensions, as well as a list of dimensions and facts. A red circle highlights the **KEY\_FACTOR** and **IMPORTANCE** dimensions under the **CD\_BUYERS** fact. A callout box points to these dimensions with the text: "Oracle Data Mining results available to Oracle BI EE administrators".
- Physical Panel:** This panel shows the physical data sources and their mappings. A red circle highlights the **CD\_BUYERS** fact and its dimensions, including **KEY\_FACTOR** and **IMPORTANCE**. A callout box points to these dimensions with the text: "Oracle Data Mining results available to Oracle BI EE administrators".

For Help, press F1

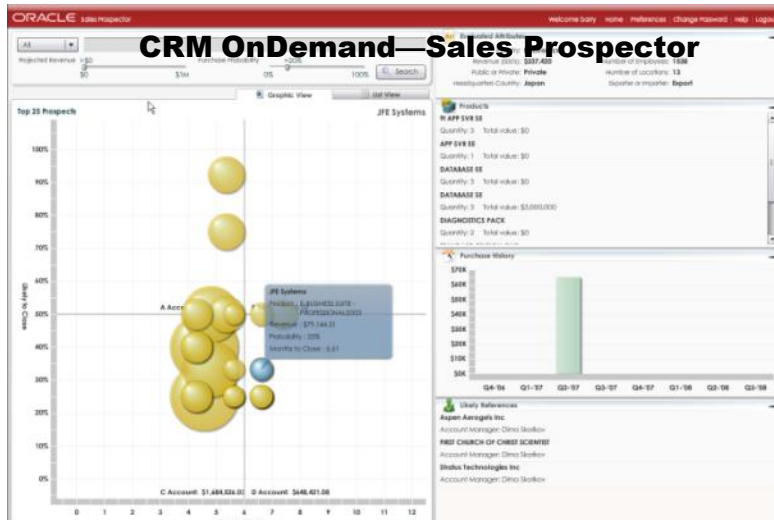
# Example

## Better Information for OBI EE Reports and Dashboards



# Predictive Analytics Applications

## Powered by Oracle Data Mining (Partial List as of March 2010)



### Oracle Communications Data Model

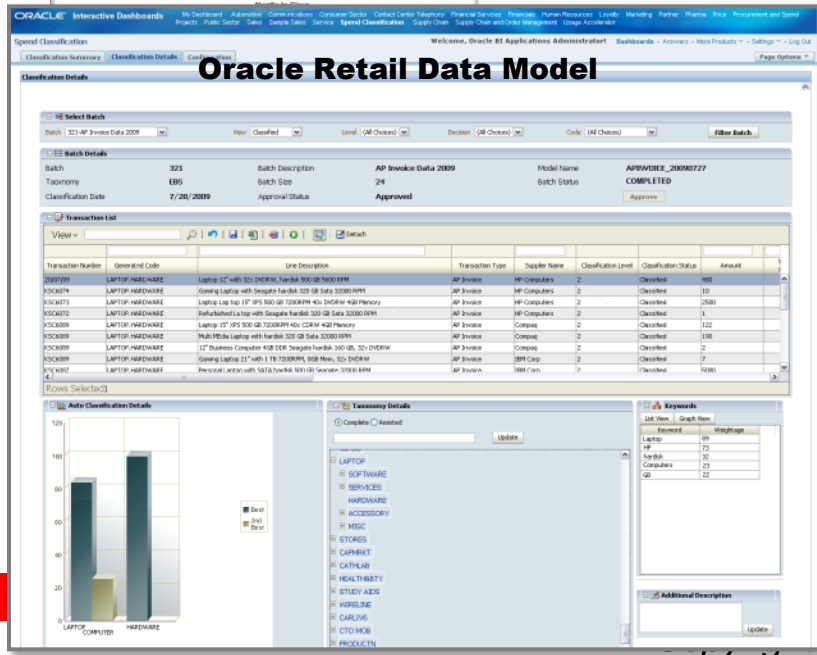
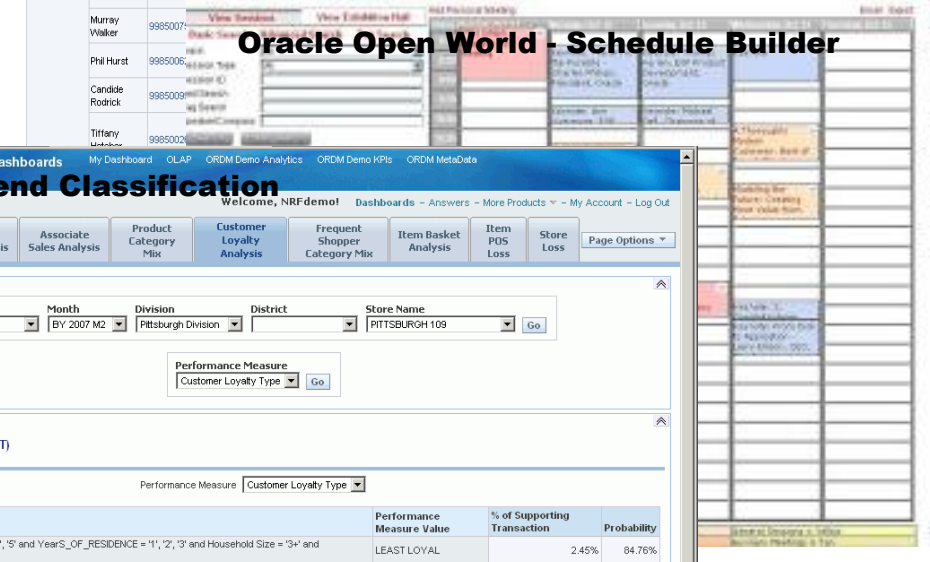
Churn by Customer Segment

Customer Segment Name is equal to Segment\_3

Customer Segment	Customer Name	Cell Phone No	Contract Value	Month Revenue	Debt Value	LTV Band	LTV Value	LTV Months	ARPU Band	Churn Indicator	Sentiment	Churn Probability	Customer Segment Key
	Chloe White	9985005370	\$0.00		\$222.00		\$65,000.00	10			▲+	56	101
	Delora Walker	9985009300	\$0.00		\$130.00		\$85,000.00	18			▲+	30	101
	Max Gerber	9985006161	\$3,000.00	\$2,500.00	\$222.00		\$79,000.00	17			▲+	39	101
	Olen Christian	9985008393	\$0.00		\$130.00		\$59,000.00	13		●	▼-	82	101
	Mason Murray	9985007979	\$9,000.00	\$7,500.00	\$70.00		\$56,000.00	21		●	▼-	94	101
	Deb Coe	9985007379	\$3,000.00	\$2,500.00	\$130.00		\$89,000.00	10			▲+		101

Probability of Churning is very high

Probability of Churning is very high



### Spend Classification

Welcome, NRFdemo! Dashboards - Answers - More Products - My Account - Log Out

Associate Loss Analysis Associate Sales Analysis Product Category Mix Customer Loyalty Analysis Frequent Shopper Category Mix Item Basket Analysis Item POS Loss Store Loss

Year: 2007 Month: BY 2007 M2 Division: Pittsburgh Division District: Store Name: PITTSBURGH 109 Go

Performance Measure: Customer Loyalty Type Go

Performance Measure: Customer Loyalty Type

file	Performance Measure Value	% of Supporting Transaction	Probability
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '1', '2', '3' and Household Size = '3+' and IDENCE = '1'	LEAST LOYAL	2.45%	84.76%
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '1', '2', '3' and Household Size = '3+' and IDENCE = '1' and Marital Status = 'DIVORCED', 'SEPARATED'	LEAST LOYAL	0.13%	100.00%
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '1', '2', '3' and Household Size = '3+' and IDENCE = '1' and Marital Status = 'MARRIED', 'SINGLE'	LEAST LOYAL	1.60%	87.85%
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '4', '5' and Household Size = '3+' and IDENCE = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '4', '5' and Household Size = '3+' and Marital Status = 'MARRIED', 'SINGLE'	PRETTY LOYAL	10.93%	83.17%
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '4', '5' and Household Size = '3+' and Marital Status = 'MARRIED', 'SINGLE'	PRETTY LOYAL	7.39%	81.58%
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '4', '5' and Household Size = 'LESS THAN 3'	MARGINALLY LOYAL	9.60%	86.29%
once = '1', '2', '3', '4', '5' and Year's_OF_RESIDENCE = '4', '5' and Household Size = 'LESS THAN 3' and 'MARRIED', 'SINGLE'	MARGINALLY LOYAL	6.52%	84.17%
once = '10', '8', '7', '6', '5' and Household Size = '3+' and Marital Status = 'MARRIED', 'SINGLE'	MOST LOYAL	25.01%	88.28%
once = '10', '8', '7', '6', '5' and Household Size = '3+' and Marital Status = 'MARRIED', 'SINGLE'	MOST LOYAL	17.11%	86.10%
once = '10', '8', '7', '6', '5' and Household Size = 'LESS THAN 3'	PRETTY LOYAL	26.20%	80.88%

ORACLE

# Fusion HCM Predictive Analytics

ORACLE® Search All [ ] [ ] [ ] [ ] You are logged in as Les Preferences Help

Navigator Recent Items Favorites Tags Watchlist

Welcome General Accounting Manager Resources

My Organization

Organization Chart Hierarchy Grid Table Number of Levels to be Displayed 2

General Employment Availability Budget Compensation Performance

Actions View Filter

Name	Job Title	Phone	Email	Id	Predicted Risk	
					Individual	Group
Leslie Hann	Finance Director	(813) 419-0861	lhann@vision.com	12345	High	High
Anna Pascal	Senior Accountant Manager	(986) 819-8861	apascal@vision.com	23456	Medium	Medium
George White	Finance Manager	(542) 465-6424	gwhite@vision.com	74342	High	High
Jason Blake	Senior Manager	(251) 331-1816	jblake@vision.com	67890	High	High
Pat Miller	Accounting Manager	(793) 465-3470	pmiller@vision.com	34567	High	High
Stella Hahn	Manager	(700) 796-6338	shahn@vision.com	90123	Medium	Medium

Predicted Worker Performance and Attrition

View My Directs' Organization Full Analysis

Average Team Prediction for My Direct Reports' Organization

☒ Show names

Team Name	Total Number of Team Members	Average Predicted Attrition	Average Predicted Performance	Prediction Details
Anna Pascal	0	High	High	
Team: George White	10	Medium	Medium	
Team: Jason Blake	6	Medium	Medium	
Team: Pat Miller	15	Medium	Medium	
Stella Hahn	0	Medium	Medium	

Predicted Attrition

High

Medium

Low

Stella Hahn

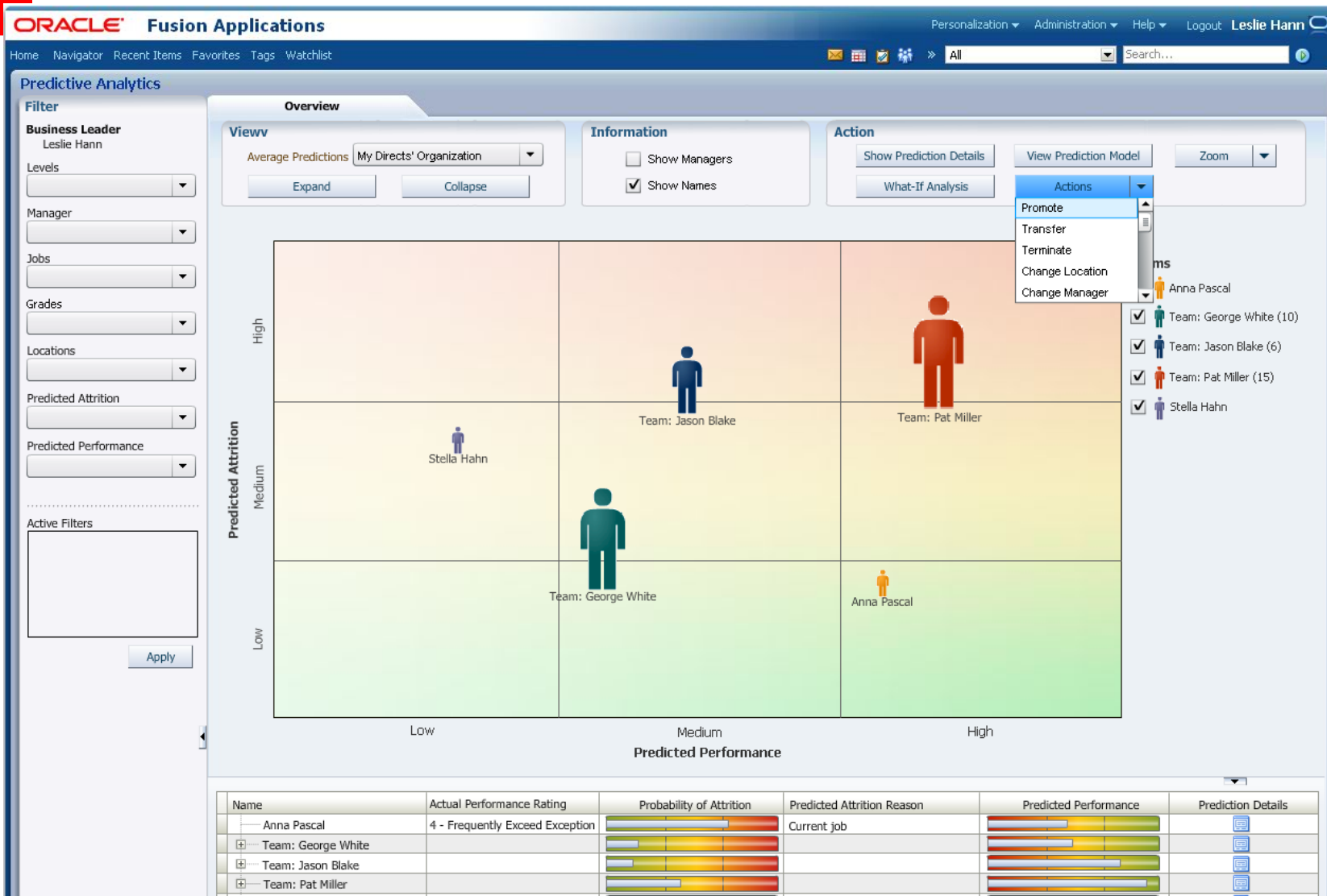
Anna Pascal

Team: Pat Miller (15)

Team: George White (10)

Team: Jason Blake (6)

# Fusion HCM Predictive Analytics







# Getting Started

SOFTWARE.  
HARDWARE.  
**COMPLETE.**

ORACLE®

# Getting Started

- Oracle Data Miner Cue Cards—part of client install
- Oracle By Example Online Learning on OTN



## Using Oracle Data Miner for Oracle Database 11g Release 2

### Purpose

This tutorial covers the use of Oracle Data Miner to perform data mining against Oracle Database 11g Release 2. The Oracle Data Miner graphical user interface (GUI), The Oracle Data Miner GUI is included as an extension of Oracle SQL Developer.

Oracle SQL Developer is a free graphical tool for database development. With SQL Developer, you can browse database statements. Starting with SQL Developer, version 3.0, you can also access the Oracle Data Miner GUI, which provides a graphical user interface for data mining.

DISCLAIMER: This tutorial has been developed with pre-production software, and is not available for external audit.

### Time to Complete

Approximately 30 mins.

### Overview

Data mining is the process of extracting useful information from masses of data by extracting patterns and trends.

- ❑ Predict individual behavior, for example, the customers likely to respond to a promotional offer or the customer likely to churn.
- ❑ Find profiles of targeted people or items (Classification using Decision Trees)
- ❑ Find natural segments or clusters (Clustering)
- ❑ Identify factors more associated with a target attribute (Attribute Importance)
- ❑ Find co-occurring events or purchases (Associations, sometimes known as Market Basket Analysis)
- ❑ Find fraudulent or rare events (Anomaly Detection)

The phases of solving a business problem using Oracle Data Mining are as follows:

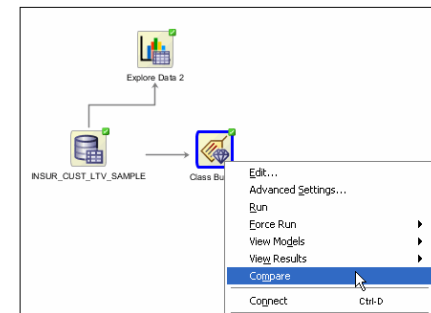
1. Problem Definition in Terms of Data Mining and Business Goals
2. Data Acquisition and Preparation
3. Building and Evaluation of Models
4. Deployment

### Compare the Models

After you build and train the selected models, you can view and evaluate the results for all of the models in a comparative format.

Follow these steps:

1. Right-click the build node and select **Compare** from the menu.



Results: A Class Build display tab opens, showing a graphical comparison of the four models, as shown here:

