

ORACLE®

ORACLE®

# Finding Gold in Your Data Warehouse: Oracle Advanced Analytics

Charlie Berger

Sr. Director Product Management, Data Mining and  
Advanced Analytics

[charlie.berger@oracle.com](mailto:charlie.berger@oracle.com)

[www.twitter.com/CharlieDataMine](http://www.twitter.com/CharlieDataMine)

Hardware and Software  
Engineered to Work Together

ORACLE  
OPEN  
WORLD

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.

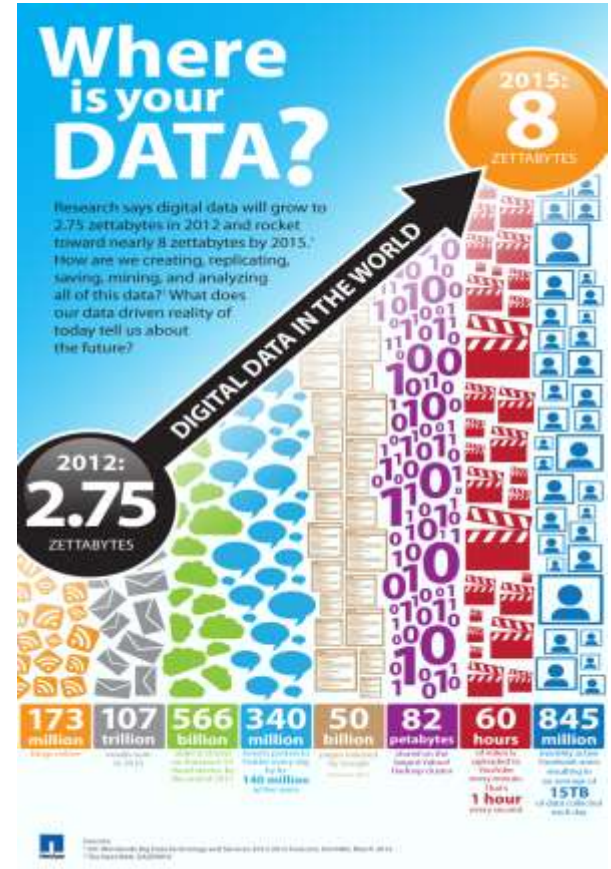
The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

# Program Agenda

- Big Data & Big Data Analytics
- Oracle Advanced Analytics
- Details
- Demos & Applications
- Pointers & Summary

# “Big Data” is Growing

- 383+ Million Twitter accounts
- 835+ Million Facebook subscribers
- 1.2+ Billion Mobile Web users
- Machine and sensor data
- Over 6 million OnStar subscribers



# “Big Data”= Structured & Unstructured Data

Exhibit 25: Structured Data Example

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	\$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	\$13.89	Yes	183	Both
M6	1	5g	6	Button	\$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	\$11.98	No	298	Phillips
M10	1.5	6g	10	Round	\$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	\$18.26	No	998	Flat
M14	2	7g	14	Round	\$21.19	No	235	Phillips
M16	2	8g	16	Button	\$23.57	Yes	292	Both
M18	2.1	8g	18	Button	\$25.87	No	664	Both
M20	2.4	8g	20	Pan	\$29.09	Yes	486	Both
M24	2.55	9g	24	Round	\$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	\$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	\$41.32	No	434	Both
M50	4.5	15g	50	Pan	\$44.72	No	740	Flat

Structured data from applications.

Exhibit 26: Quasi-Structured Data Example

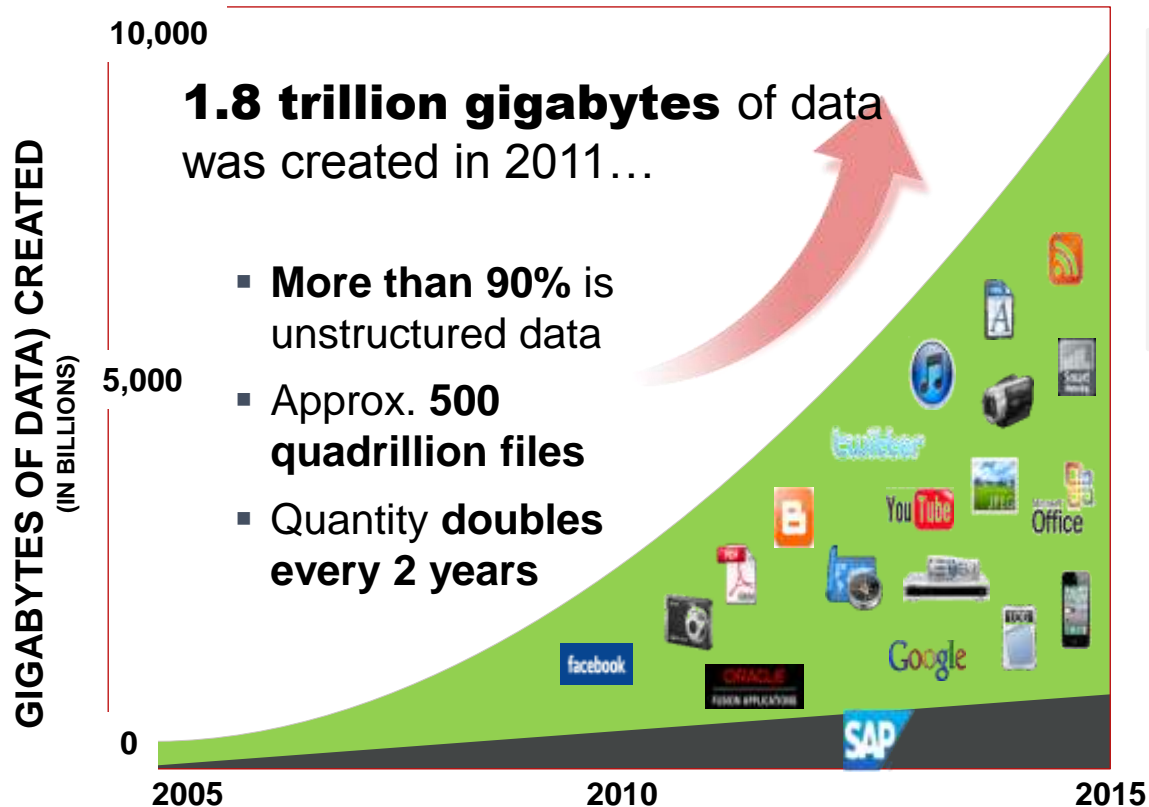
```

ex081018.log - Notepad
File Edit Format View Help
4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.NET+
HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;
07.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+K
-80 - 24.207.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+
Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPa
943130030 80 - 24.207.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.
4+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.
/2.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+I
I2MPH1K6LH5y+dyUC4467WSEF43DNWU1E1_08KUC24E1HMT03H1E_AUVD0G1HFBV3F20W8S12C+8.
I2MPH1K6LH5y+dyUC4467WSEF43DNWU1E1_08KUC24E1HMT03H1E1HWHU0145RNL45DCBLW1Cdq
/7.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CL
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+.
-80 - 24.207.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+
-80 - 24.207.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+
-80 - 24.207.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+
HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.43
.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1
HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.432
5 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4
.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CL
- 24.207.152.165 HTTP/1.1 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath
 Mozilla/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPa
tergencewebx2fDefault.aspx?3fTabid%3d34 80 - 24.207.152.165 HTTP/1.1 Mozilla/4.0+(c
/4.0+(compat|ble|;+MSIE+6.0;+Windows+NT+5.1;+SVL;+.NET+CLR+1.1.4322;+InfoPath.2;+
  
```

Semi-structured “Big Data” from social media and logs, sensors, feeds, etc.



# “Big Data” → “Big Data Analytics”



*“There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing.”*

- Google CEO Eric Schmidt

## Requires capability to rapidly:

- ✓ **Collect and integrate data**
- ✓ **Understand data & their relationships**
- ✓ **Respond and take action**

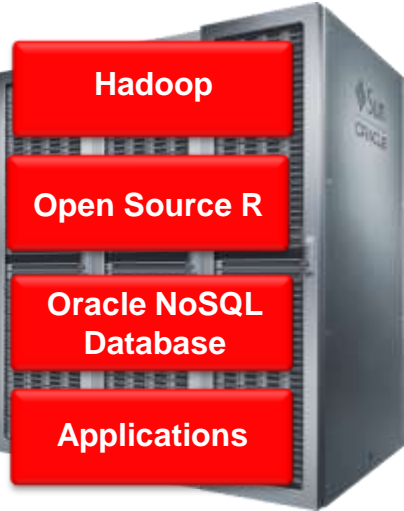
ORACLE

# Oracle Big Data Platform



## Oracle Big Data Appliance

Optimized for Hadoop, R, and NoSQL Processing

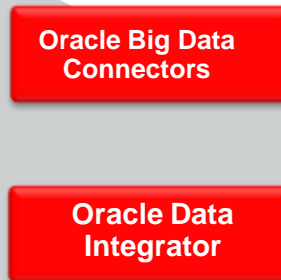


Stream

Acquire

Organize

## Oracle Big Data Connectors



## Oracle Exadata

“System of Record”  
Optimized for DW/OLTP



Discover & Analyze

## Oracle Exalytics

Optimized for Analytics & In-Memory Workloads



“Without *proper analysis*, it's just data; ...not useful actionable information ...something that you can exploit today ...something that your competitor may *not* have yet discovered.”

## Charlie Berger

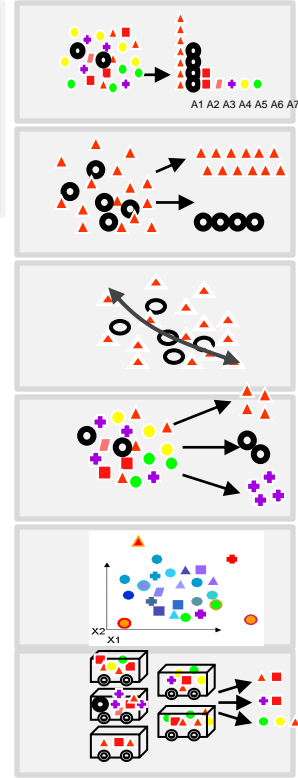
Sr. Director, Product Management, Oracle Data Mining and  
Advanced Analytics  
Oracle Corporation



# What is Data Mining?

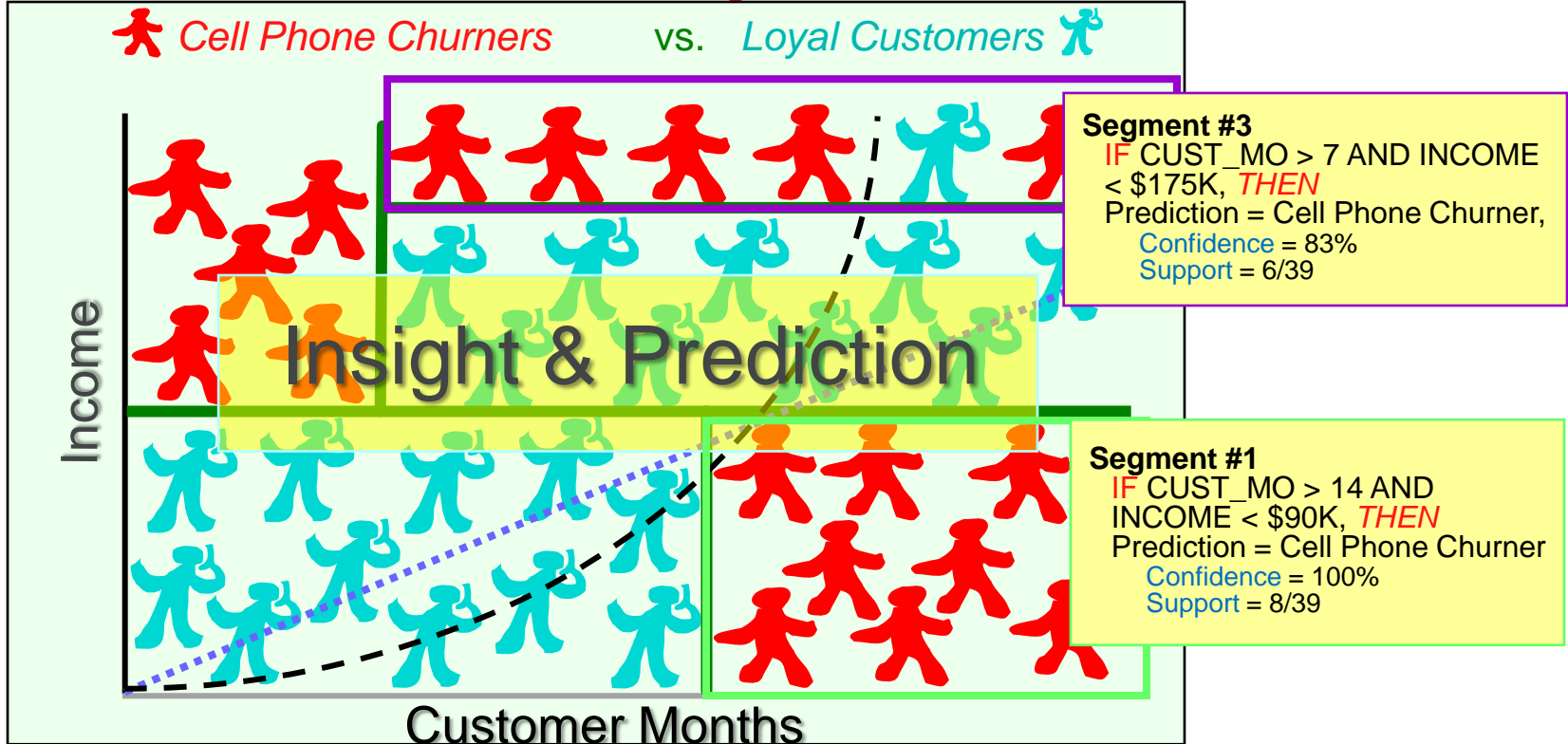
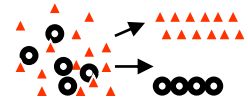
*Automatically* sifting through large amounts of data to find previously hidden patterns, discover valuable new insights and make predictions

- Identify most important factor (*Attribute Importance*)
- Predict customer behavior (*Classification*)
- Predict or estimate a value (*Regression*)
- Find profiles of targeted people or items (*Decision Trees*)
- Segment a population (*Clustering*)
- Find fraudulent or “rare events” (*Anomaly Detection*)
- Determine co-occurring items in a “baskets” (*Associations*)



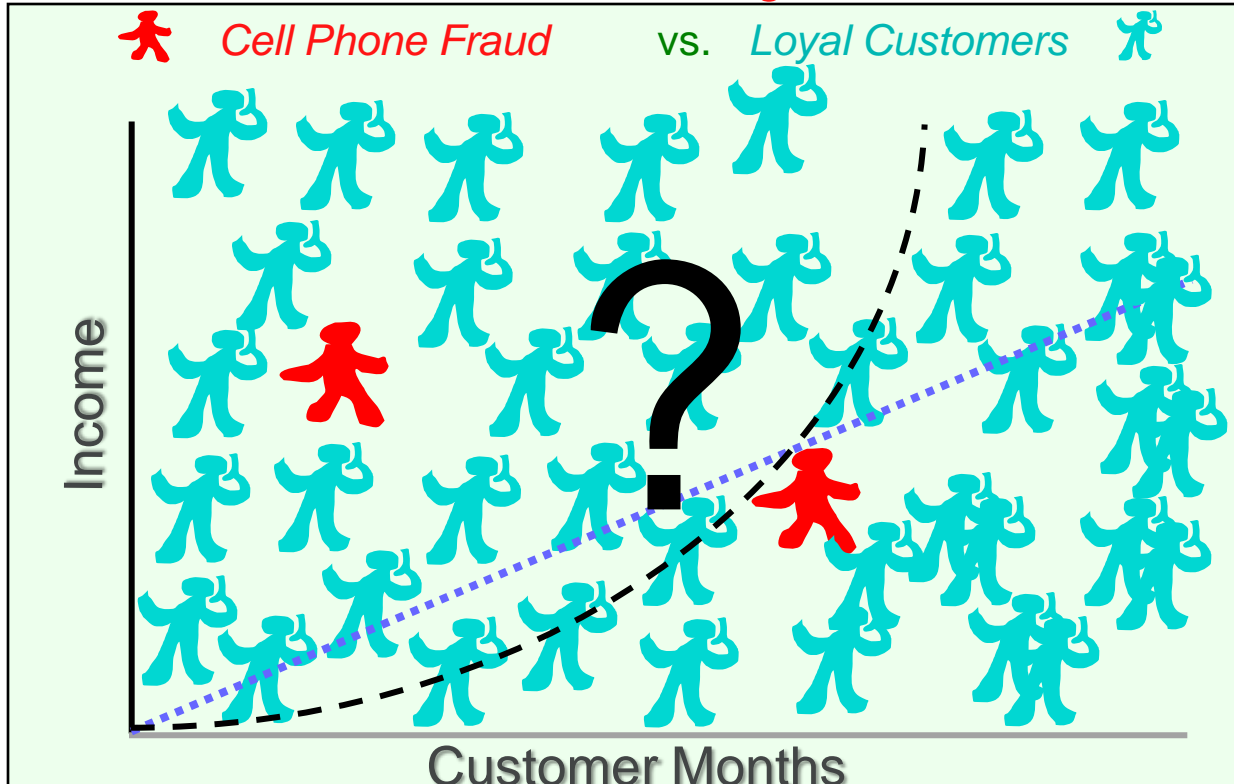
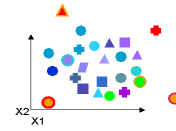
# Data Mining Provides

Better Information, Valuable Insights and Predictions



# Data Mining Provides

Better Information, Valuable Insights and Predictions

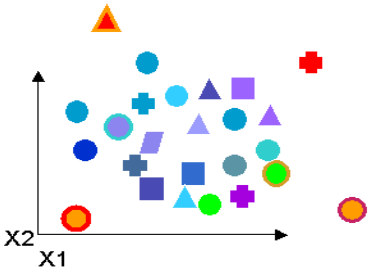


# Finding Needles in Haystacks

- Haystacks are usually **BIG**
- Needles are typically *small* and **rare**



# Look for What is “*Different*”



# A Real Fraud Example

My credit card statement—**Can you see the fraud?**



Total purchases exceeds  
time period average

May 22	1:14 PM	FOOD	Monaco Café	\$127.38
May 22	7:32 PM	WINE	Wine Bistro	\$28.00
...				
June 14	2:05 PM	MISC	Mobil Mart	<u>\$75.00</u>
June 14	2:06 PM	MISC	Mobil Mart	<u>\$75.00</u>
June 15	11:48 AM	MISC	Mobil Mart	<u>\$75.00</u>
June 15	11:49 AM	MISC	Mobil Mart	<u>\$75.00</u>
May 28	6:31 PM	WINE	Acton Shop	\$31.00
May 29	8:39 PM	FOOD	Crossroads	\$128.14
June 16	11:48 AM	MISC	Mobil Mart	<u>\$75.00</u>
June 16	11:49 AM	MISC	Mobil Mart	<u>\$75.00</u>

Gas Station?

Monaco?

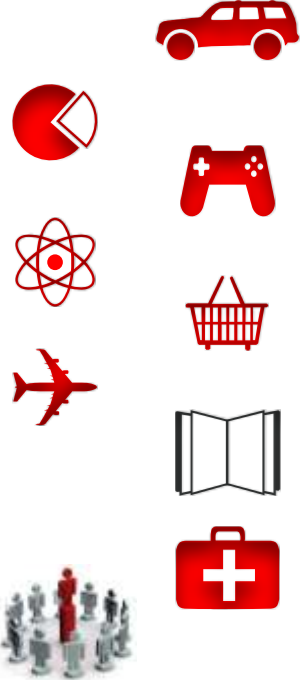
Pairs of  
\$75?

All same \$75 amount?

# Data Mining & Predictive Analytics

## Typical Use Cases

- Targeting the right customer with the right offer
- How is a customer likely to respond to an offer?
- Finding the most profitable growth opportunities
- Finding and preventing customer churn
- Maximizing cross-business impact
- Security and suspicious activity detection
- Understanding sentiments in customer conversations
- Reducing medical errors & improving quality of health
- Understanding influencers in social networks



# Oracle Advanced Analytics Details

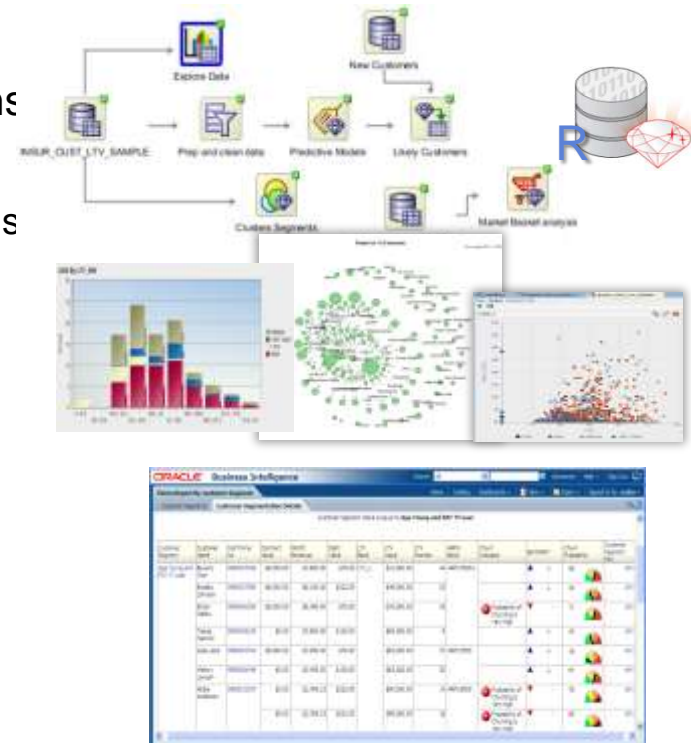


ORACLE

# Oracle Advanced Analytics Option

## Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

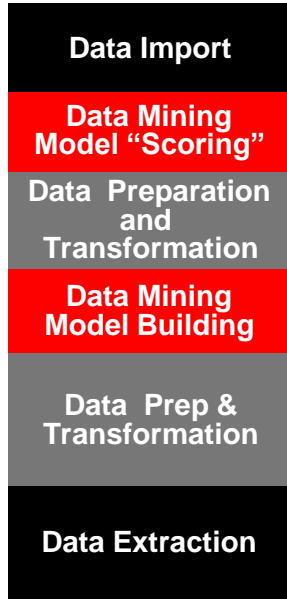
- **Powerful**
  - Combination of in-database data mining algorithms and open source R algorithms
  - Accessible via SQL, PL/SQL, R and database APIs
  - Scalable, parallel in-database execution
- **Easy to Use**
  - Range of GUI and IDE options for business users to data scientists
- **Enterprise-wide**
  - Integrated feature of the Oracle Database
  - Seamless support for enterprise analytical applications and BI environments



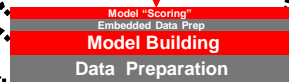
# Oracle Advanced Analytics Value Proposition



## Traditional Analytics



## Oracle Advanced Analytics



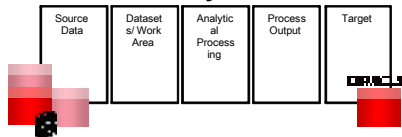
**Savings**

### Value Proposition

- Fastest path from data to insights
  - Fastest analytical development
  - Fastest in-database scoring engine on the planet
- Flexible deployment options for analytics
- Lowest TCO by eliminating data duplication
- Secure, Scalable and Manageable

- Data remains in the Database
- Data preparation for analytics is automated
- Scalable distributed-parallel implementation of machine learning techniques in the database
- Scalable implementation of R programming language in-database
- Flexible interface options – SQL, R, IDE, GUI
- Fastest and most Flexible analytic deployment options
- Can import 3<sup>rd</sup> party models

### Hours, Days or Weeks



### Secs, Mins or Hours



# Oracle Advanced Analytics

## Target Audiences

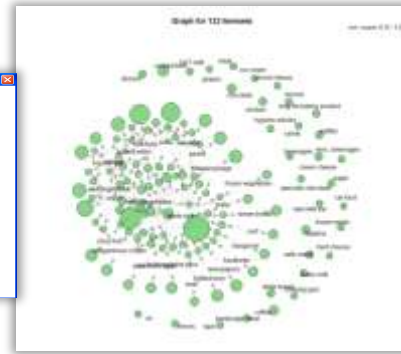
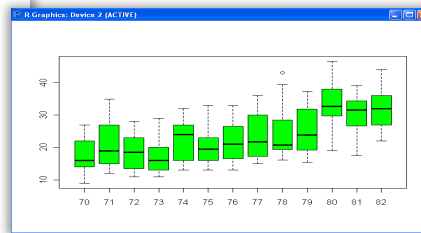
### ▪ “Information Consumers”

- CEOs, CMOs, CFOs, CIOs, VPs, Directors/Mgrs of lines of operations, etc.



### ▪ “Information Producers”

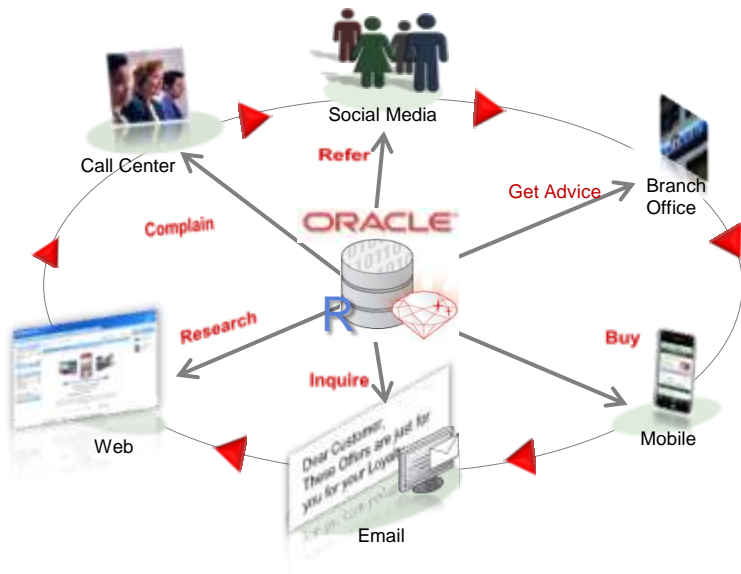
- Data analysts, Marketing analysts, Business Managers, Statisticians, Data Scientists, DBAs, Application Developers



# Oracle Advanced Analytics

## More Details

- Advanced Analytic scoring tasks can now be executed **in batch or real-time** in **CRM systems** where customers can be contacted via **targeted call-center/websites/e-mail offers**



OAA's predictive models can “power” any real-time environments and deliver personalized predictions, clustering, recommendations, etc.

**Likelihood to respond:**

Query Result

SQL | All Rows Fetched: 1 in 0 seconds

PREDICTION_PROB...
0.8382936507936...

# Oracle Advanced Analytics

## More Details

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

```
Select prediction_probability(CLAS_DT_1_1, 'Yes'
  USING 7800 as bank_funds, 125 as checking_amount, 20 as
  credit_balance, 55 as age, 'Married' as marital_status,
  250 as MONEY_MONTHLY_OVERDRAWN, 1 as house_ownership)
from dual;
```



**Likelihood to respond:**

Query Result

SQL | All Rows Fetched: 1 in 0 seconds

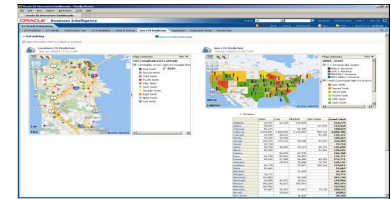
PREDICTION_PROB...
0.8382936507936...



# Enabling Predictive Applications

## Example Applications Using Oracle Advanced Analytics

- **Human Capital Management**
  - **Predictive Workforce**—employee turnover and performance prediction and “What if?” analysis
- **CRM**
  - **Sales Prediction Engine**--prediction of sales opportunities, what to sell, amount, timing, etc.
- **Supply Chain Management**
  - **Spend Classification**—real-time flagging of possible noncompliance in expense submissions
- **Identity Management**
  - **Oracle Adaptive Access Manager**—real-time security and fraud analytics
- **Retail Analytics**
  - **Oracle Retail Customer Analytics**—”shopping cart analysis” and next best offers
- **Customer Support**
  - **Predictive Incident Monitoring (PIM)** Customer Service offering for Database customers
- **Manufacturing**
  - Response surface modeling in chip design
- **Predictive capabilities in Oracle Industry Data Models**
  - **Communications Data Model** implements churn prediction, segmentation, profiling, etc.
  - **Retail Data Model** implements loyalty and market basket analysis
  - **Airline Data Model** implements analysis frequent flyers, loyalty, etc.



# Oracle Communications Industry Data Model

## Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

ORACLE Business Intelligence

Search: All Advanced Help Sign Out

Home Catalog Dashboards New Open Signed In As ocidm

Churn Report By Customer Segment

Customer Segments Customer Segmentation Details

OAA's clustering and predictions available in-DB for OBIEE

Customer segments

Customer Segment	Customer Name	Cell Phone No	Contract Value	Month Revenue	Debt Value	LTV Band	LTV Value	Churn Probability	Customer Segment Key	Community Role	Community Size	Churner Ratio in Community	Avg Revenue of Community
Age Young and PAY TV user	Beverly Wan	9985007046	\$18,000.00	\$15,600.00	\$140.00	LTV_1	\$41,000.00	39	104	LOCAL	3	0.00%	\$1.00
	Bradley Johnson	9985007589	\$18,000.00	\$16,200.00	\$444.00		\$49,000.00	45	104	PASSIVE	3	0.00%	\$0.00
	Ethan Haeley	9985006289	\$18,000.00	\$16,800.00	\$140.00		\$34,000.00	71	104	LOCAL	4	0.00%	\$2.33
	Gale Lazar	9985003794	\$18,000.00	\$14,000.00	\$140.00		\$82,000.00	16	104	PASSIVE	7	2.00%	\$8.75
	Bernard Vaughn	9985005144	\$6,000.00	\$5,478.26	\$260.00		\$85,000.00	19	104	LOCAL	4	1.00%	\$3.00
	Bertha Lucca	9985002105	\$6,000.00	\$5,355.56	\$444.00		\$56,000.00	76	104				\$3.50
	Bett Webber	9985000594	\$6,000.00	\$5,538.46	\$380.00		\$76,000.00	16	104				\$5.00

Segment Name

Age Young and PAY TV user

(All Column Values)

- Age Young and PAY TV user
- Bad phone number and Low usage
- Family User, High Revenue
- High and insensitive to Loyalty Program
- High value Organizational Customer
- High value and use loyalty program
- Low Revenue

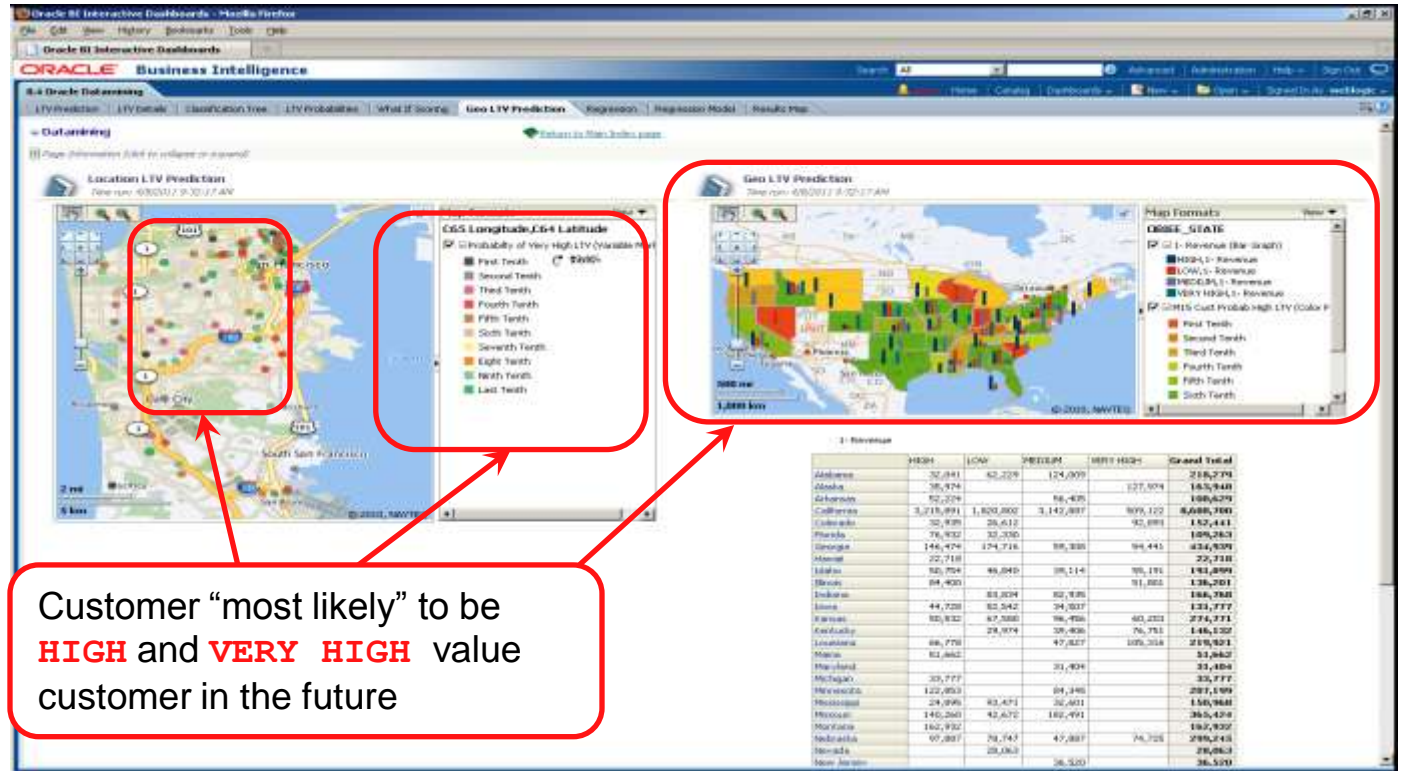
Probability of Churning is very high

Probability of Churning is very high

# Integrated Business Intelligence

Integrate a range of in-DB SQL & R Predictive Analytics & Graphics

- In-database construction of predictive models that predict customer behavior
- OBIEE's integrated spatial mapping shows where



# Fusion HCM Predictive Analytics

## Built-in Predictive Analytics

The screenshot displays the 'Predicted Worker Performance and Attrition' interface. On the left, a 2x2 matrix plots 'Predicted Attrition' (Low to High) against 'Predicted Performance' (Low to High). Data points include Stella Hahn, Anna Pascal, Team: Pat Miller (15), Team: George White (10), and Team: Jason Blake (6). A red box highlights the 'Team: Pat Miller (15)' data point, with an arrow pointing to the 'Prediction Details' panel on the right.

The 'Prediction Details: Team: Pat Miller' panel shows summary statistics: Manager Pat Miller, Average Predicted Performance 78% (High), Average Predicted Attrition 77% (High), and Total Number of Workers in Group 15. Below this is a table with two tabs: 'Predicted Attrition' and 'Predicted Performance'. The 'Predicted Performance' tab is active, showing a table of team members with their predicted attrition levels and top contributing factors.

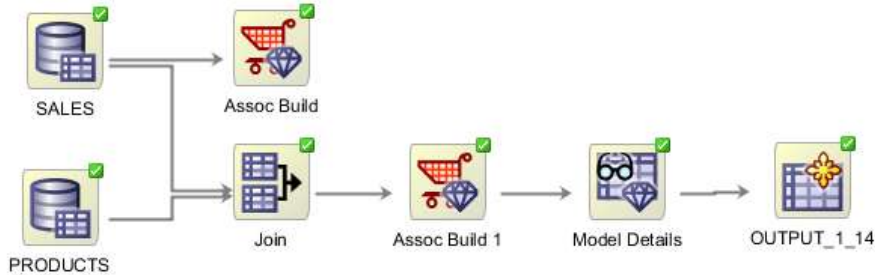
Team Name	Predicted Attrition	Top Most Contributing Factor
Carmelo Frink	High	Most Recent Salary Change
James Jones	Medium	Time Spent in Current Position
Julie Lee	Medium	Time Spent in Current Position
Sam Smith	Medium	Time Spent in Current Position
Joe Jones	Medium	Most Recent Salary Change
Justin Rico	Medium	Amount of Sickness
Paula Gupton	Medium	Most Recent Salary Change
Edward Espinoza	Medium	Time Spent in Current Position
Lisa Trahan	Medium	Home Country
Laura Wang	Medium	Time Spent in Current Position
Scott Henderson	Medium	Most Recent Salary Change
Patrick Bottom	Medium	Marital Status
Michael Hernandez	Medium	Time Spent in Current Position
Pat Miller	Medium	Most Recent Salary Change
		Working Hours

A red box highlights the table content, with an arrow pointing to a callout box at the bottom of the slide.

Oracle Advanced Analytics factory-installed predictive analytics show employees likely to leave, top reasons, expected performance and real-time "What if?" analysis

# Retail GBU

## Market Basket Analysis



Market Basket Analysis to identify co-occurring items found in “baskets” and potential product bundles

Rules | Itemsets | Settings

Sort by: Lift Descending

Fetch Size: 1,000

Rule Content: Name, Subname

Rules: 1,000 out of 20,988

ID	Antecedent	Consequent	Lift	Confidence(%)	Support(%)
17313	PROD_ID.137 AND PROD_ID.143 AND PROD_ID.138	PROD_ID.144	29.1235	70.8661	1.0689
17849	PROD_ID.138 AND PROD_ID.143 AND PROD_ID.139	PROD_ID.144	28.7389	69.9304	1.0528
17673	PROD_ID.137 AND PROD_ID.143 AND PROD_ID.142	PROD_ID.144	28.3021	68.8675	1.1003
20557	PROD_ID.139 AND PROD_ID.143 AND PROD_ID.142	PROD_ID.144	27.8527	67.7739	1.0975
17314	PROD_ID.137 AND PROD_ID.144 AND PROD_ID.138	PROD_ID.143	27.8265	77.5469	1.0689
18071	PROD_ID.138 AND PROD_ID.146 AND PROD_ID.144	PROD_ID.143	27.5925	76.8947	1.0207
17942	PROD_ID.138 AND PROD_ID.144 AND PROD_ID.140	PROD_ID.143	27.5082	76.6599	1.0486
18075	PROD_ID.138 AND PROD_ID.148 AND PROD_ID.144	PROD_ID.143	27.4991	76.6346	1.0563

Rule Details:

ID: 17313

IF  
 PROD\_ID.137 AND  
 PROD\_ID.143 AND  
 PROD\_ID.138

THEN  
 PROD\_ID.144

Lift	29.1235
Confidence(%)	70.8661
Support(%)	1.0689
Antecedent Support(%)	2.4333
Consequent Support(%)	1.5083
Item Count	3

# Big Data Analytics in Retail

- Leverage all customer touch-point information
  - Consider each customer's demographics and past and recent POS behavior
  - POS data—shifting “market basket” items
  - Identify customer segments (“Country Squires”, “Green”, “New Empty Nests”)
  
- Deploy real-time predictive models
  - 1:1 Marketing—treat each customer as an individual relationship
  - Look for opportunities to combine multiple touch points
  - Geo-location provides opportunity for site specific recommendations
  - Changing consumption, provide opportunities for cross-selling/up-selling



# Big Data Analytics in Financial Services

## ■ 360° View of Customer

- Integrate silos of multi-business CRM data within large corporations
- Combine data from multiple sources: investments, retail banking, mortgages
- Gain 360° perspective of all touch points with a customer
- Develop “best” customer profiles and sell them the right product at the right time



## ■ Identify and combat fraud

- Real-time fraud detection
- Transactional data combine with demographic data
- Monitor velocity of recent purchases and checks written vs. hist. averages
- Flag transactions and individuals that appear “different” from normal behavior



# Big Data Analytics High Performance Operations

- Learn from manufacturing, warranty, service data
  - Devices report back product's performance: analyze part failure correlations and patterns
  - Identify new strategies for improved product design and service plans
  - Increase product uptime, performance and quality



- Characterize and understand all performance scenarios
  - Streaming data from multiple sensors, weather, water, etc.
  - Clustering and response modeling to optimize each scenario

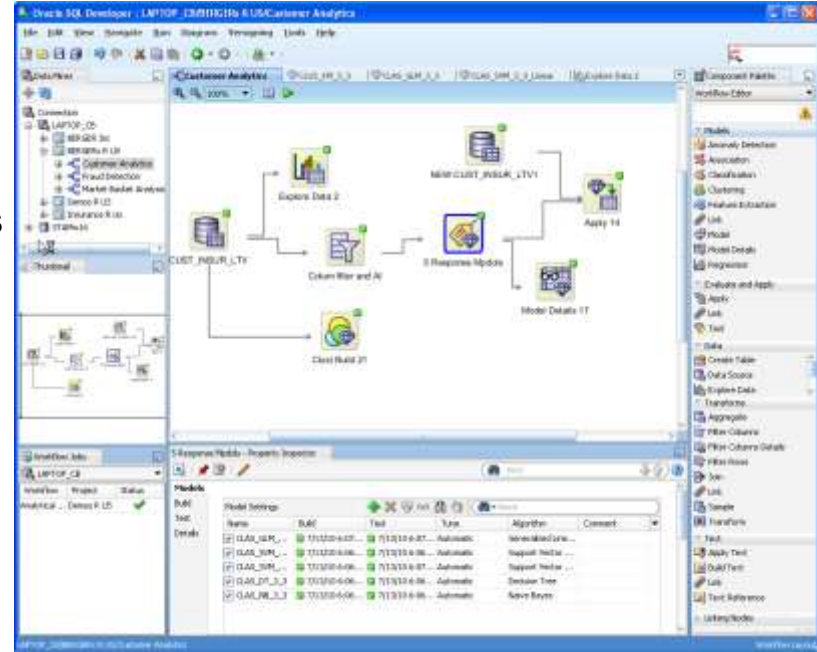
*“The USA holds 250 sensors to collect raw data: pressure sensors on the wing; angle sensors on the adjustable trailing edge of the wing sail .... But collecting data was only the beginning. BMW ORACLE Racing also had to manage that data, analyze it, and present useful results. The team turned to Oracle Data Mining in Oracle Database 11g.”*



# Oracle Data Miner GUI

## SQL Developer 3.2 Extension—Free OTN Download

- Easy to Use
  - Oracle Data Miner GUI for data analysts
  - Explore data—discover new insights
  - “Work flow” paradigm for analytical methodologies
- Powerful
  - Multiple algorithms & data transformations
  - Runs 100% in-DB
  - Build, evaluate and apply data mining models
- Automate and Deploy
  - Generate and deploy SQL scripts for automation
  - Share analytical workflows



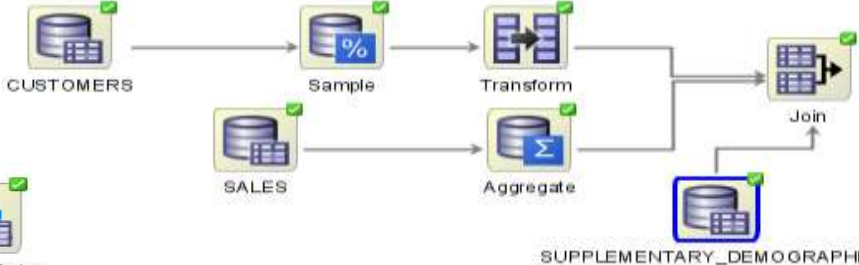
# Oracle Data Miner GUI

## Oracle Data Miner Nodes — *Partial List*

### Tables and Views



### Transformations



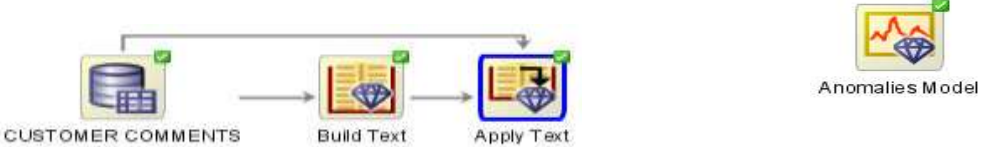
### Explore Data



### Modeling



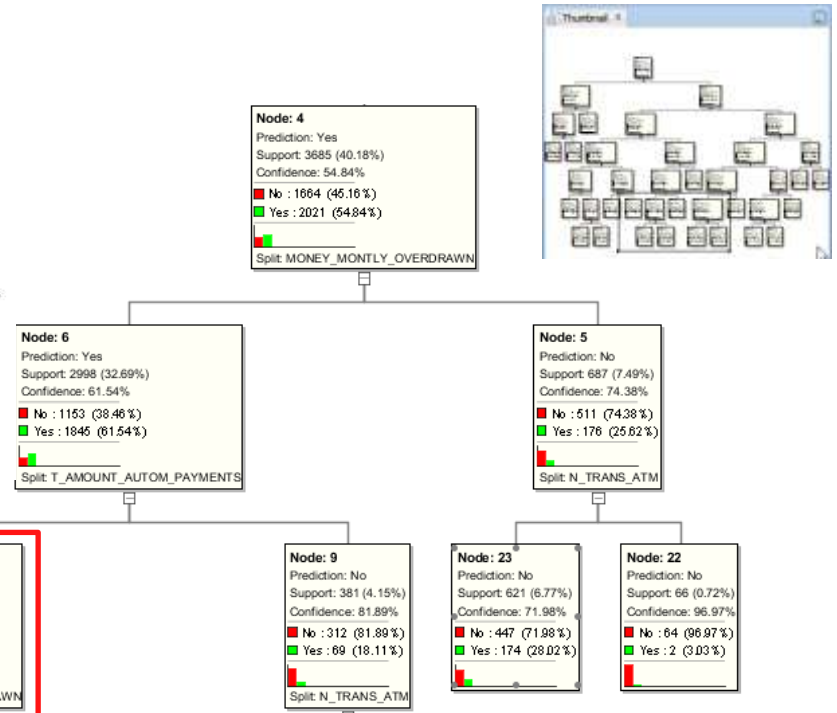
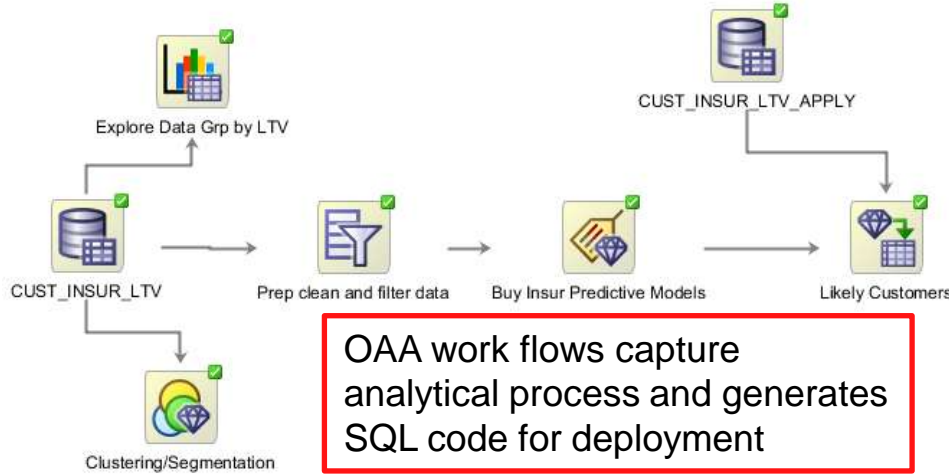
### Text



# Insurance



## Identify “Likely Insurance Buyers” and their Profiles



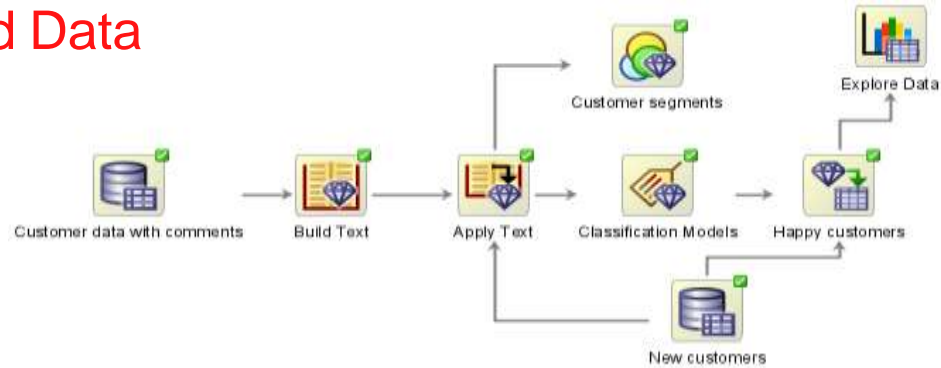
```
Rule Surrogates Target Values
If BANK_FUNDS > 225.5
And CHECKING_AMOUNT <= 207.5
And MONEY_MONTHLY_OVERDRAWN > 53.115
And T_AMOUNT_AUTOM_PAYMENTS > 8283.5
And N_TRANS_ATM > 6.5
Then Yes
```

```
Node: 7
Prediction: Yes
Support: 2617 (28.53%)
Confidence: 67.86%
No : 841 (32.14%)
Yes : 1776 (67.86%)
Split: MONEY_MONTHLY_OVERDRAWN
```

# Oracle Advanced Analytics

## Data Mining Unstructured Data

- Mines unstructured i.e. “text” data
- Include text and comments in models
- Cluster and classify documents
- Oracle Text used to preprocess unstructured text



Output Sample
CUST_ID
101509
101510
101511
101512
101513
101514

COMMENTS
Shopping at your store is a hassle. I rarely shop there and usually forget to bring your new loyalty card and hence never get the items at the sale price. Can a store manager look up my account on-line?

COMMENTS_TOK	
Name	Frequency
NEW	1
PRICE	1
RARELY	1
SALE	1
SHOP	1
SHOPPING	1
STORE	2
UP	1
USUALLY	1



# 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;
```

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

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

## Automated Monthly “Application”! *Just*

*add:*

Create

View CLAIMS2\_30

As

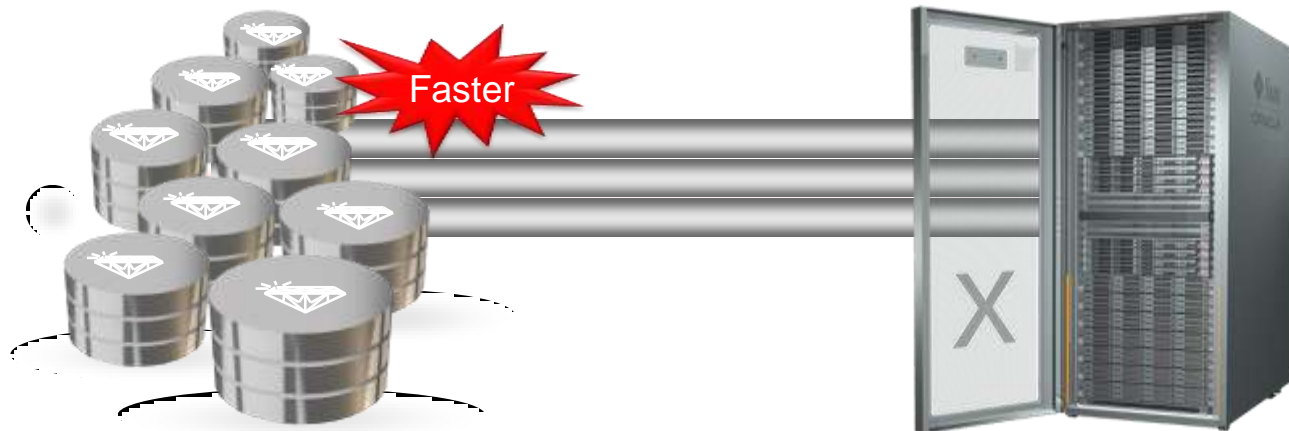
Select \* from CLAIMS2

Where mydate > SYSDATE – 30

```
-- 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;
```

# Exadata + Data Mining 11g Release 2

## Data Mining Model "Scoring" Pushed to Storage



- SQL predicates and OAA 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;
```

**Exadata "smart scan" SQL function scoring**



# Oracle Advanced Analytics

## SQL Data Mining Algorithms

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

# Oracle Advanced Analytics

## SQL Statistics and SQL Analytics



- **Ranking functions**
  - rank, dense\_rank, cume\_dist, percent\_rank, ntile
- **Window Aggregate functions**  
(moving & 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

Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition and Enterprise Edition

ORACLE

# Independent Samples T-Test

(Pooled Variances)



- Query compares the mean of AMOUNT\_SOLD between MEN and WOMEN within CUST\_INCOME\_LEVEL ranges. Returns observed t value and its related two-sided significance

```
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;
```

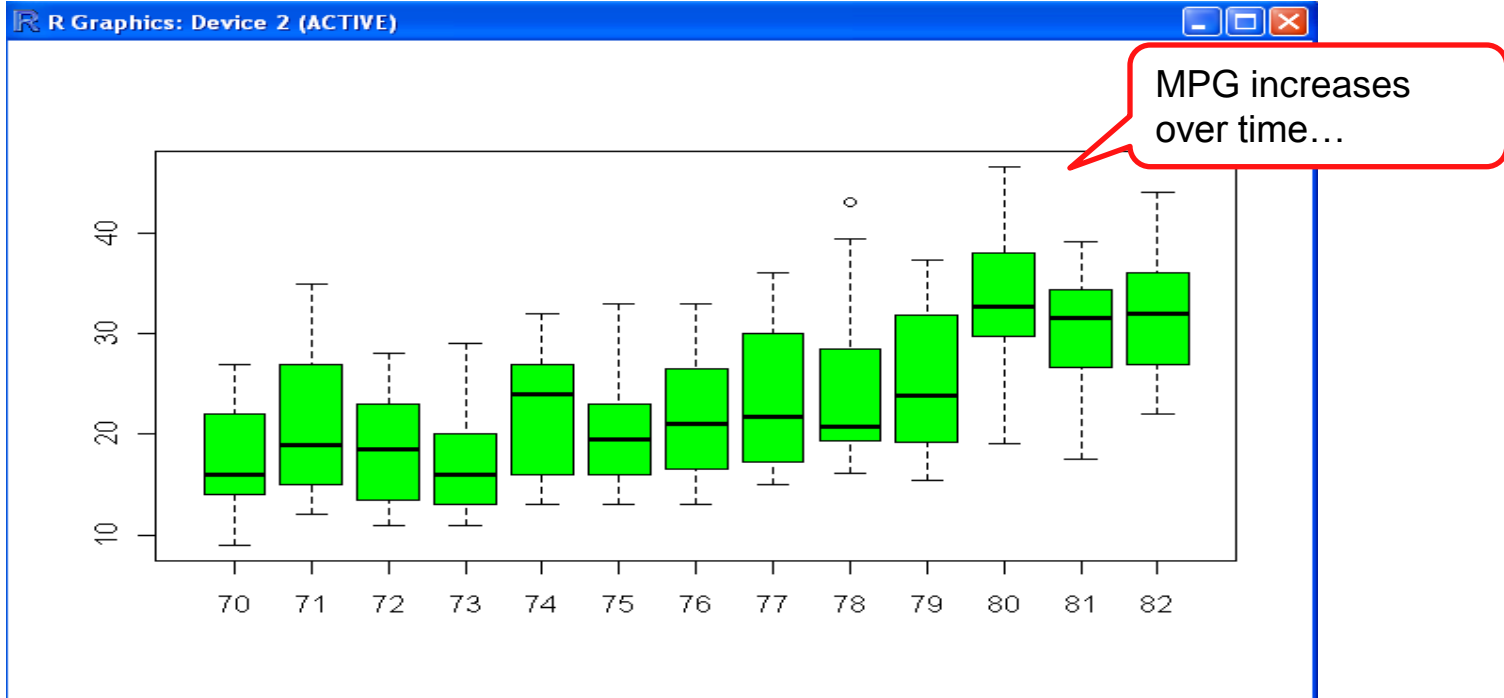
SQL Plus

ORACLE

# Oracle Advanced Analytics

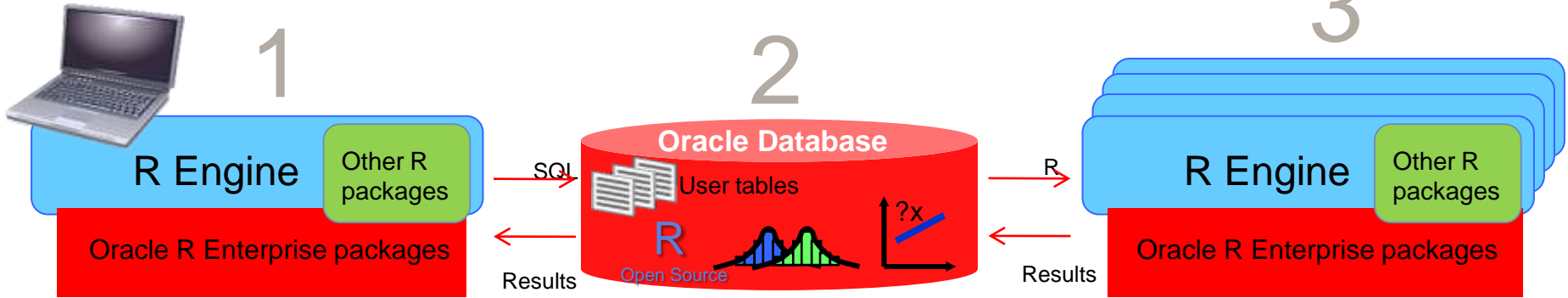
## R Graphics Direct Access to Database Data

```
R> boxplot(split(CARSTATS$mpg, CARSTATS$model.year), col = "green")
```



# Oracle Advanced Analytics

## R Enterprise Compute Engines



### User R Engine on desktop

- R-SQL Transparency Framework intercepts R functions for scalable in-database execution
- Function intercept for data transforms, statistical functions and advanced analytics
- Interactive display of graphical results and flow control as in standard R
- Submit entire R scripts for execution by database

### Database Compute Engine

- Scale to large datasets
- 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 Advanced Analytics Example

## Use of All 3 ORE Engines Within 1 R Script

The following example illustrates use of all 3 engines from within 1 R script.

```
nary = relational_table_1;
nary = relational_table_2;
m3 <- merge(nary, nary, by="ID", all.y=T) /* Join */
summary(m3) /* Summary */
tabulate(m3)
aggregate(m3$ID, by=list(age=AGE, gen=GEN), FUN=median)
BP <- boxplot(list(age=AGE), plot=FALSE)
ore_corr(m3, 'AGE, CLASS', group.by='COUNTRY', 'EDUCATION')
filtered_set <- m3[AGE==17 & !is.na(AGE) && COUNTRY IN ('USA', 'France'), c(1,4,5)]
```

Data preparation/analysis/joins/visualization/summarization/correlation/cross tabulation -> completely shipped to database for execution.

```
x = ore.pull(filtered_set)
library(arma)
y = arma(ore.pull)
```

Local pull of filtered subset for processing by an open source package – `arma` in this case. Result of processing added to database object `m3` as a derived column -> Local R engine on user's desktop

```
m3$newcolumn = y;
library(bigm)
res <- ore.tableApply(m3[c("ARRDELAY", "DISTANCE", "newcolumn")],
  function(dat) {
    library(bigm)
    biglm(ARRDELAY ~ DISTANCE + newcolumn, dat)
  })
class(mod)
mean(res$SPRED, na.rm = TRUE)
cat(res, BP); /* object returned from the script */
```

Embedded R engine invoked to build a model on the prepared/augmented data using yet another open source package `biglm` in this case



# You Can Think of OAA 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 Advanced Analytics (SQL & R)

- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine/analyze

### • Analytical “Verbs”

- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE



# Learn More

Send [Charlie.berger@oracle.com](mailto:Charlie.berger@oracle.com) email and I'll send you my "fav links"

The screenshot shows the Oracle Learning Library interface. At the top, there's a navigation bar with 'Home', 'Advanced Search', 'Bookmarks', 'My Reviews', and 'About'. Below that, a breadcrumb trail shows 'Home > Content Details'. The main content area is titled 'Oracle Data Mining 11g Release 2 OBE Series' and includes tags for 'Application Development', 'Data Mining', and 'SQLDEV'. A list of four OBE items is displayed, each with a thumbnail, title, description, release date, duration, and average rating (indicated by stars). To the right of the list, there are two red buttons labeled 'OBE Details' with play icons.

Title	Release Date	Duration	Average Rating
Setting Up Oracle Data Miner 11g Release 2	11-MAR-201115	mins	★★★★★
Using Oracle Data Miner 11g Release 2	11-MAR-201130	mins	★★★★★
Star Schema Mining Using Oracle Data Miner	11-MAR-201130	mins	★★★★★
Text Mining Using Oracle Data Miner	11-MAR-201130	mins	★★★★★

1. Link to the [Copy of my latest OAA/ODM presentation](#).
2. Link to [Blog entry w/ YouTube-like recorded of OAA/ODM presentation and several "live" demos](#)
3. Link to [Getting Started w/ ODM blog entry](#)
4. Link to [New 2 day ODM Instructor Led OU course](#).
5. Link to [SQL Developer Days Virtual Event w/ downloadable Virtual Machine \(VM\) images of Oracle Database + ODM/ODMr and e-training for Hands on Labs](#)
6. Main [ODM on OTN](#) page
7. Main [OAA on OTN](#) page
8. Main [Oracle R Enterprise page](#) on OTN & [ORE Blog](#)



**ORACLE®**