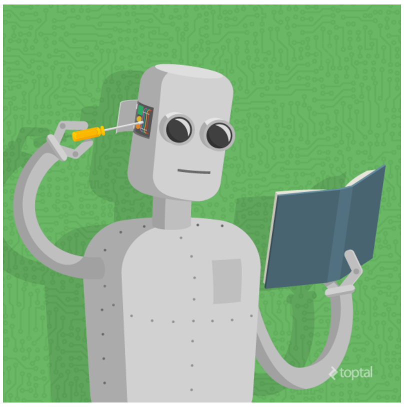


Machine Learning in ORACLE

Brendan Tierney

ORALYTICS

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

ORACLE ACE Director

- Data Warehousing since 1997
- Data Mining since 1998
- Analytics since 1993

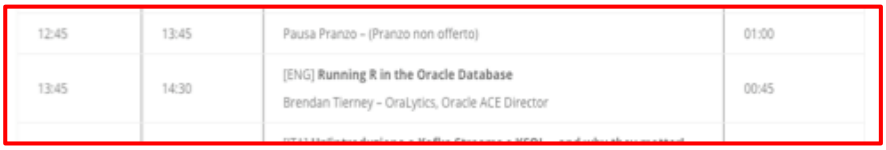
ORALYTICS

Google brendan tierney data mining

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

Topics

- Who is doing Machine Learning?
- What is Machine Learning?
- Machine Learning in Oracle
  - GUI
  - SQL & PL/SQL
  - Oracle Machine Learning (on Oracle Autonomous Data Warehouse Cloud Service)
  - Automated Machine Learning => Machine Learning for Dummies
  - R



12:45	13:45	Pausa Pranzo – (Pranzo non offerto)	01:00
13:45	14:30	[ENG] **Running R in the Oracle Database** Brendan Tierney – OraLytics, Oracle ACE Director	00:45

Who is ?

Doing Machine Learning

Oracle

R

SAS/SPSS

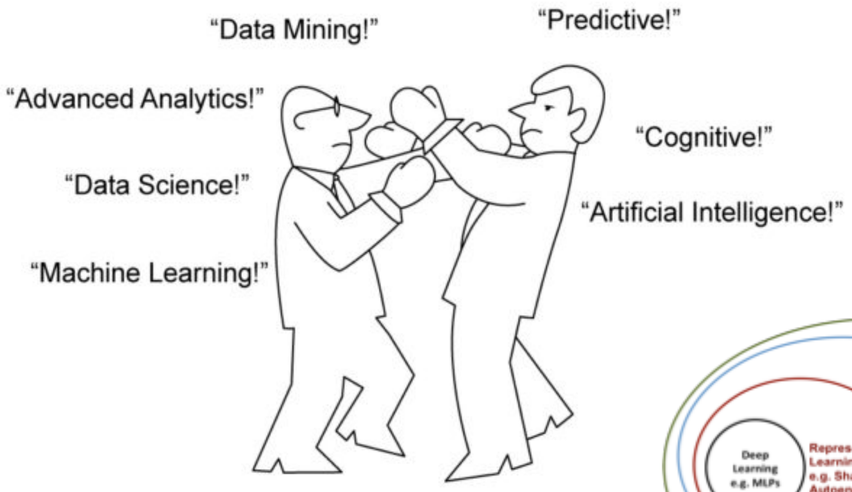
RapidMiner

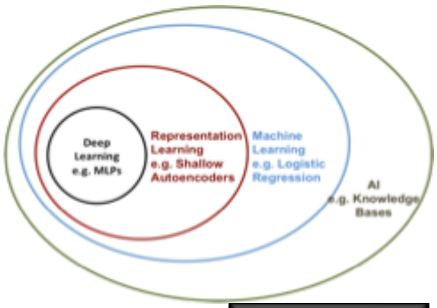
Python

Something else?

Not doing ML now?

but would like to be or Very Soon





Data Mining / Predictive Analytics / Machine Learning

- Data Mining is the non-trivial extraction of previously unknown and potentially useful information from (large collections of) data

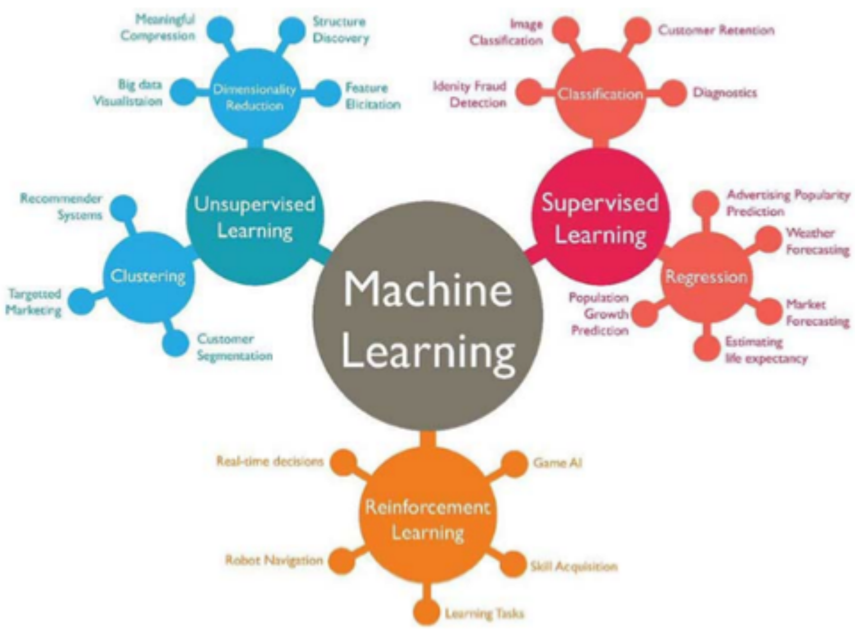
Data Mining is about explaining the past

to

predict the future.

It does not tell us some magic answer(s)

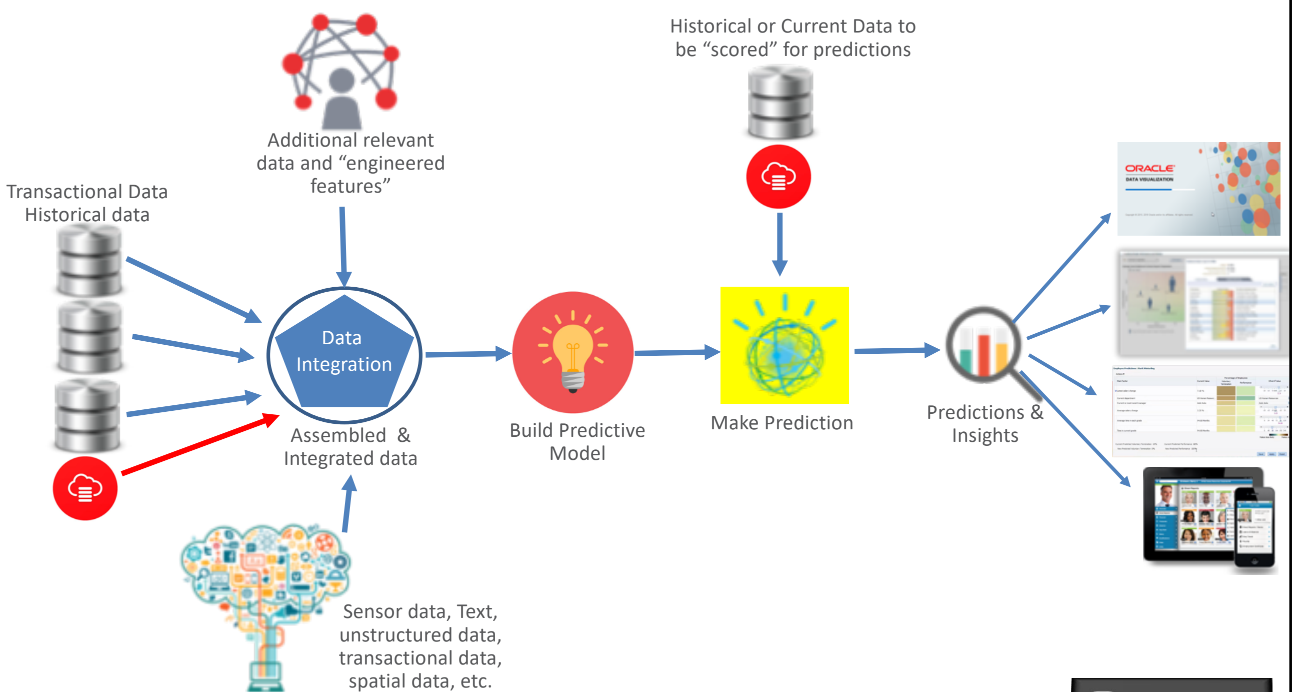
It only gives us more data (or information) which needs to be assessed to see if it is useful

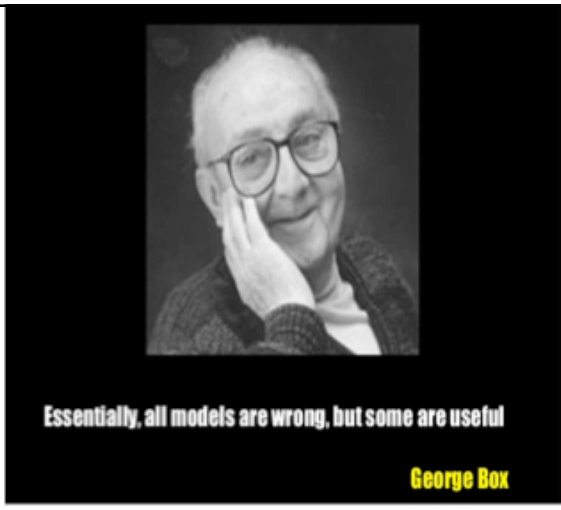


www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

Your data isn't as big as you think!



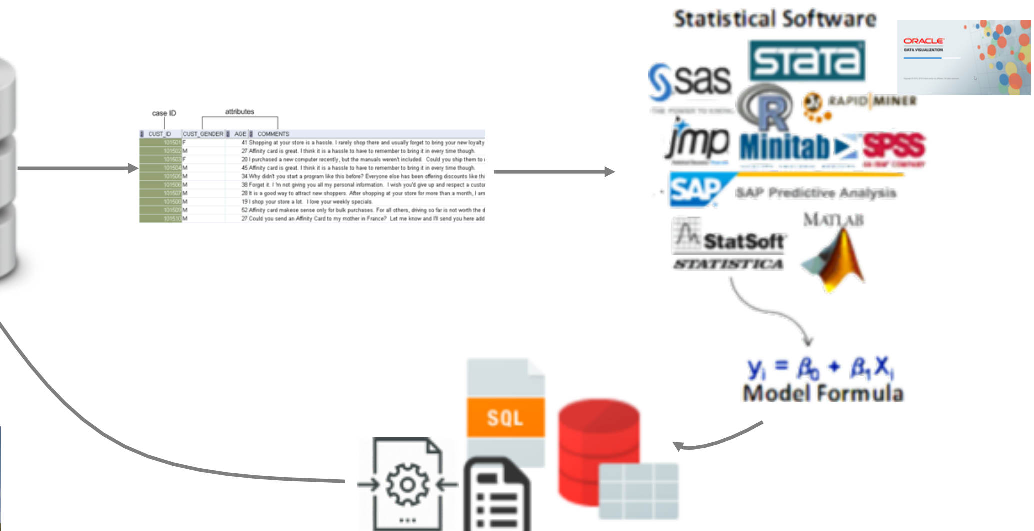


A model is a simplification or approximation of reality and hence will not reflect all of reality.

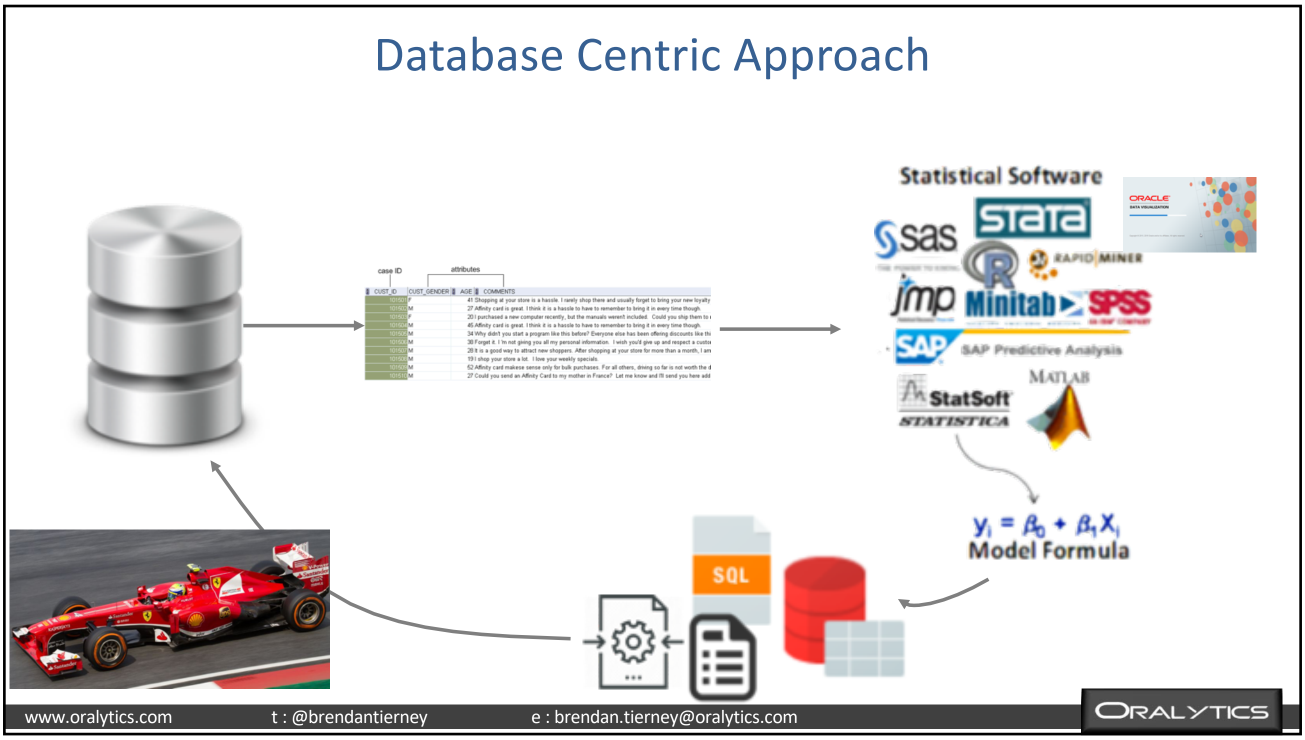
His paper was published in the Journal of the American Statistical Association, 1976
Book *Empirical Model-Building and Response Surfaces*, 1987

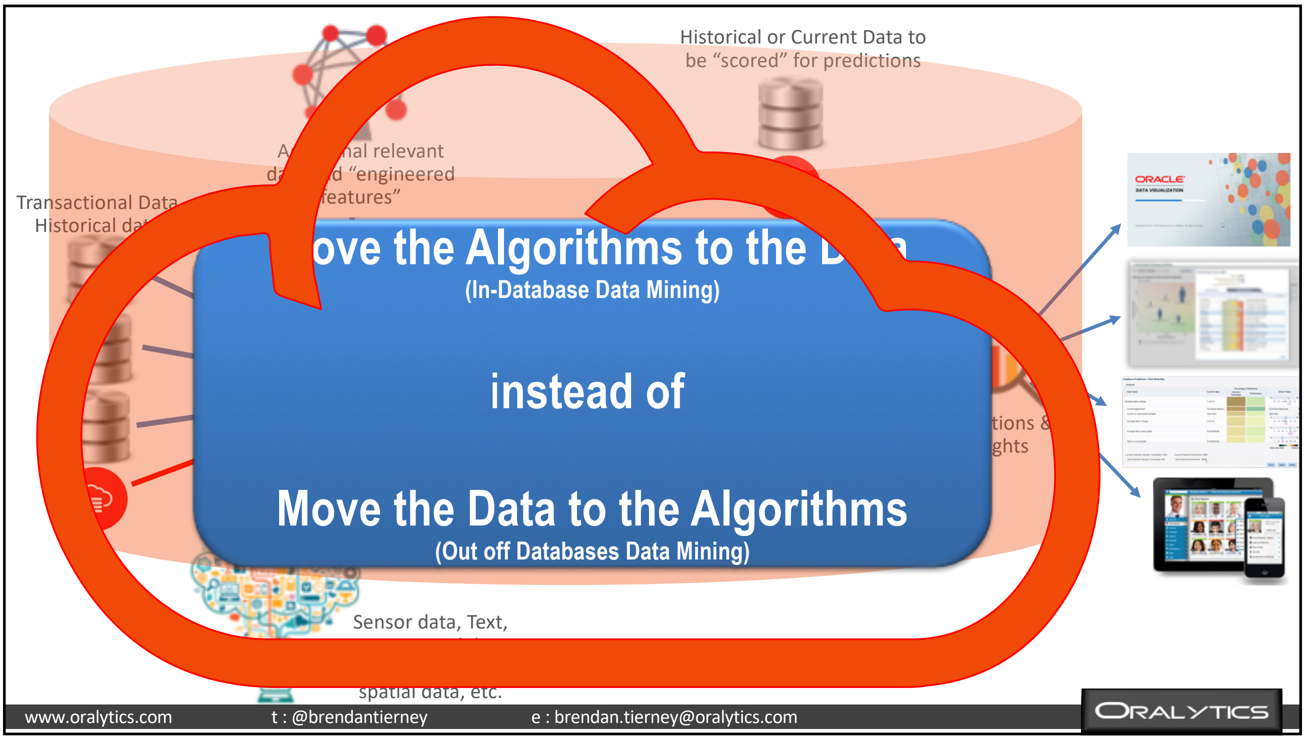
www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com

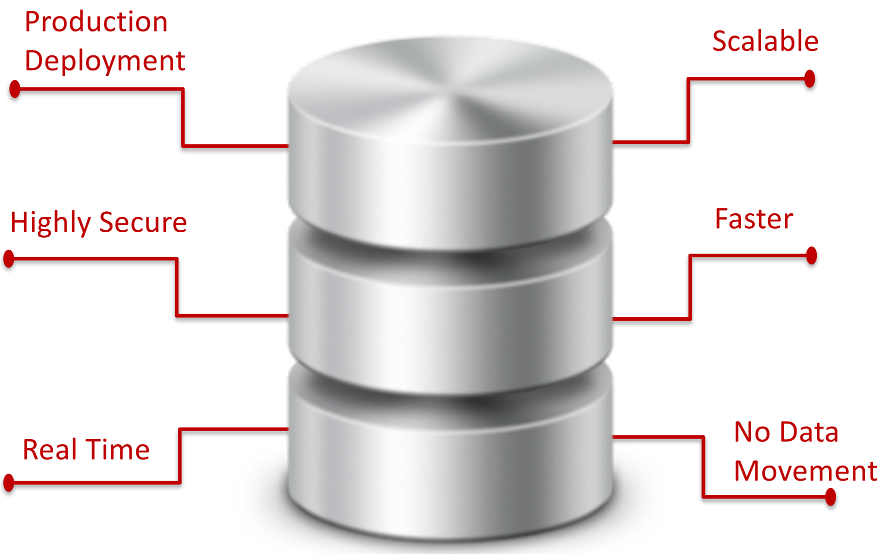
Traditional Approach



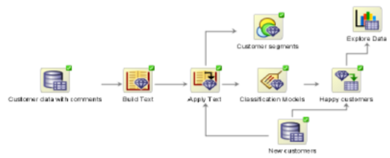
www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com


Database Centric Approach
Statistical Software
SQL
Model Formula
www.oralytics.com
t : @brendantierney
e : brendan.tierney@oralytics.com
ORALYTICS


Historical or Current Data to
be “scored” for predictions
Transactional Data
ove the Algorithms to the D
(In-Database Data Mining)
instead of
Move the Data to the Algorithms
(Out off Databases Data Mining)
Sensor data, Text,
www.oralytics.com
t : @brendantierney
e : brendan.tierney@oralytics.com
ORALYTICS



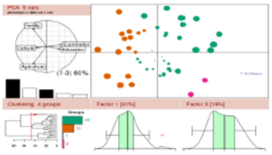
Comprehensive Advanced Analytics Platform



Oracle Data Mining

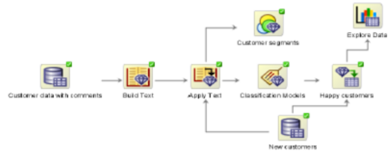
- **SQL** kernel; automated knowledge discovery inside the Database
- 18 in-database data mining algorithms
- Text mining
- Predictive analytics applications development environment
- Star schema and transactional data mining
- Exadata "scoring" of ODM models
- SQL Developer/Oracle Data Miner GUI

Comprehensive Advanced Analytics Platform



Oracle R Enterprise

- Popular open source R 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

- SQL kernel; automated knowledge discovery inside the Database
- 18 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

Statistical Functions in Oracle

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

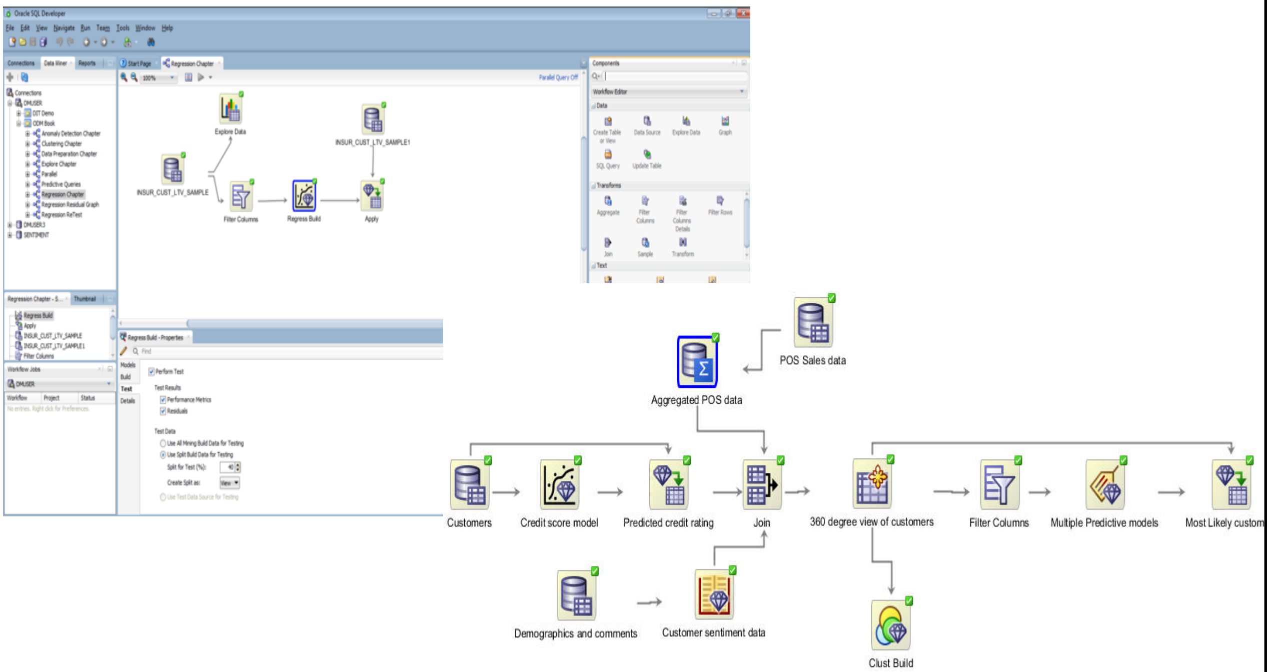
Hypothesis Testing
– Student t-test , F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov- Smirnov test, One-way ANOVA

Distribution Fitting
– Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi- Squared Test, Normal, Uniform, Weibull, Exponential

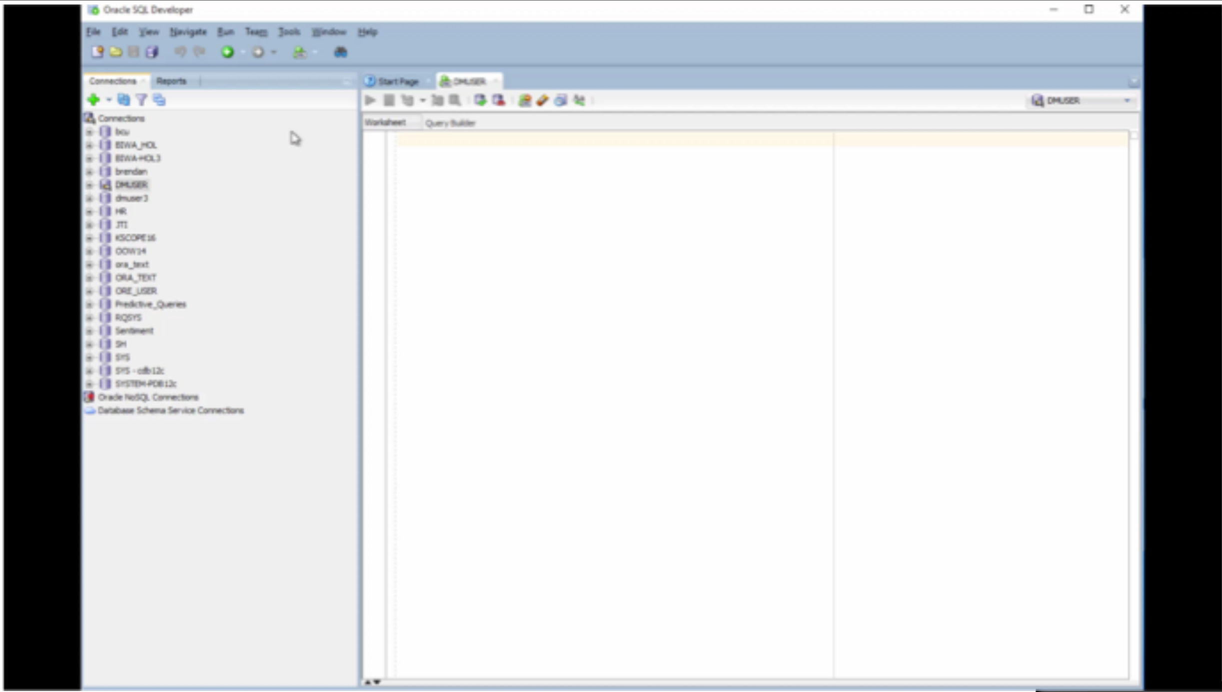
SQL

All of these are
FREE
with the Database

These are often forgotten about



You can run all the example code if you have the ODM Demo Schema created

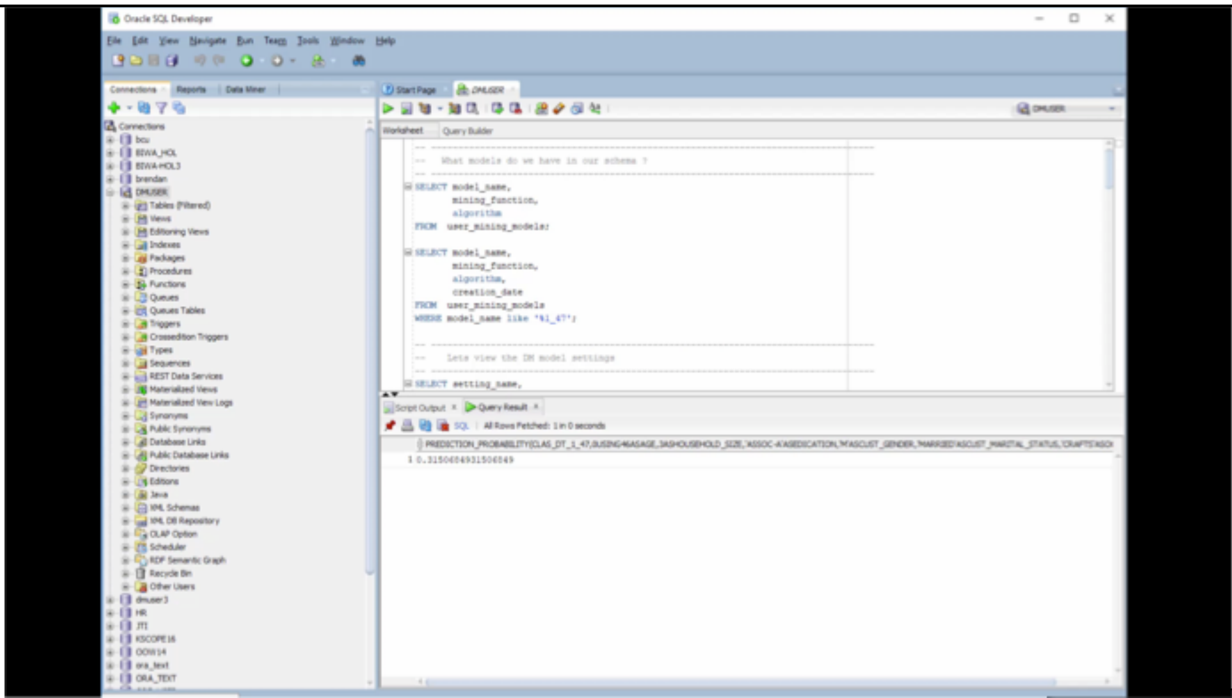


Oracle Data Mining

- PL/SQL Package
  - DBMS_DATA_MINING
  - DBMS_DATA_MINING_TRANSFORM
  - DBMS_PREDICTIVE_ANALYTICS
- 12c – Predictive Queries
  - aka Dynamic Queries
  - Transitive dynamic Data Mining models
  - Can scale to many 100+ models all in one statement
- SQL Functions
  - PREDICTION
  - PREDICTION_PROBABILITY
  - PREDICTION_BOUNDS
  - PREDICTION_COST
  - PREDICTION_DETAILS
  - PREDICTION_SET
  - CLUSTER_ID
  - CLUSTER_DETAILS
  - CLUSTER_DISTANCE
  - CLUSTER_PROBABILITY
  - CLUSTER_SET
  - FEATURE_ID
  - FEATURE_DETAILS
  - FEATURE_SET
  - FEATURE_VALUE

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com


SQL
Code Demo


Oracle SQL Developer
File Edit View Navigate Run Team Tools Window Help
Connections Reports Data Miner
Start Page DMUSER
Worksheet Query Builder
Script Output Query Result
All Rows Fetched: 1 in 0 seconds

EU GDPRs - Right to Explanation

- Article 22 - Automated individual decision-making, including profiling
- Article 21 - Right to object
- Article 13 – Information to be provided where personal data are collected from the data subject
- Article 14 - Information to be provided where personal data have not been obtained from the data subject
- Article 15 - Right of access by the data subject
- Article 17 - Right to erasure (‘right to be forgotten’)
- Article 18 - Right to restriction of processing

Article 13

(f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com

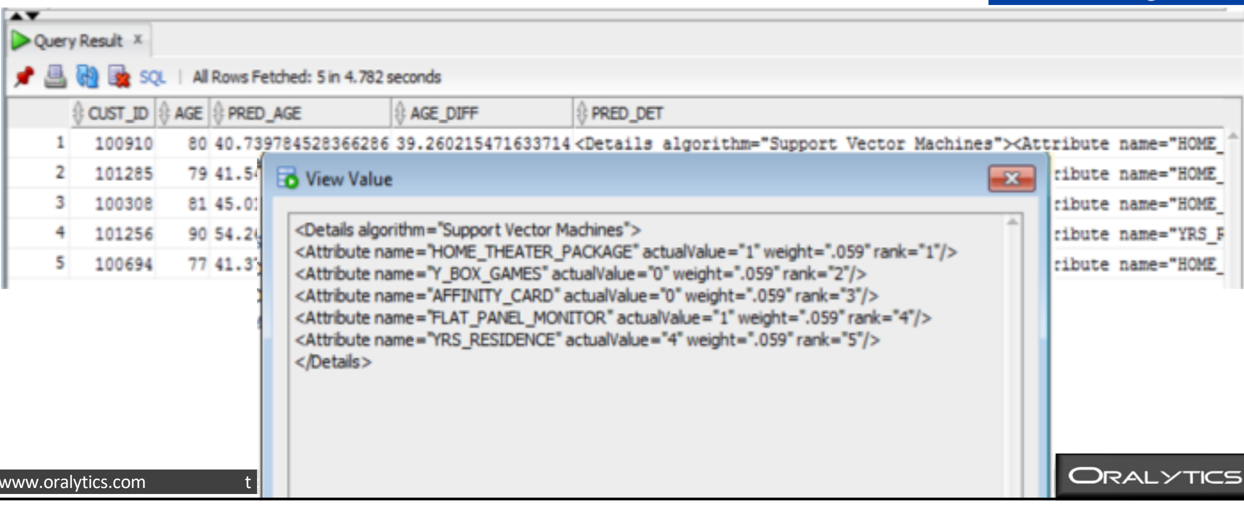
ORALYTICS

```
SELECT cust_id, age, pred_age, age-pred_age age_diff, pred_det
    FROM (SELECT cust_id, age, pred_age, pred_det,
          RANK() OVER (ORDER BY ABS(age-pred_age) DESC) rnk
          FROM (SELECT cust_id, age,
                PREDICTION(FOR age USING *) OVER ()      pred_age,
                PREDICTION_DETAILS(FOR age ABS USING *) OVER () pred_det
                FROM mining_data_apply_v))
    WHERE rnk <= 5;
```

This example dynamically identifies customers whose age is not typical for the data.

The query returns the attributes that predict or detract from a typical age.

General Data Protection Regulation



www.oralytics.com t

ORALYTICS

Companies are building products that bridge the gap between the Data Scientist and the End-User

- Machines doing Machine Learning
- Automated Machine Learning
- Autonomous Machine Learning
- Check out Predictive Queries 12c

```
select cust_id, affinity_card,
       PREDICTION( FOR to_char(affinity_card) USING *) OVER ()   pred_affinity_card
from mining_data_build_v;
```

CUST_ID	AFFINITY_CARD	PRED_AFFINITY_CARD
102226	0	1
102227	0	1
102230	1	1
102234	0	0
102235	0	0
102238	0	0
102240	0	0
102241	0	0
102243	1	1
102244	0	0
102250	1	1
102258	0	0
102260	0	0
102263	0	0
102266	1	1
102268	0	0
102271	1	1
102279	0	1

PQ to predict the AFFINITY_CARD value.

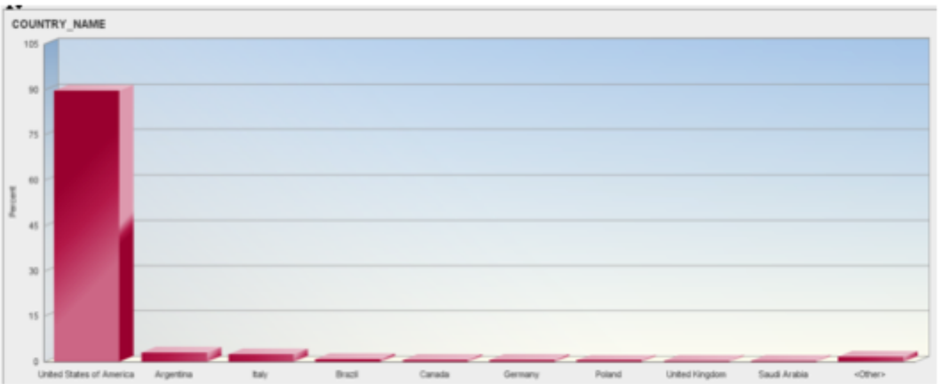
Using all the data
USING *


```
select cust_id, affinity_card,
       PREDICTION( FOR to_char(affinity_card) USING *) OVER ()   pred_affinity_card,
       PREDICTION_PROBABILITY( FOR to_char(affinity_card) USING *) OVER ()   pred_prob
from mining_data_build_v;
```

We can use other Prediction functions.

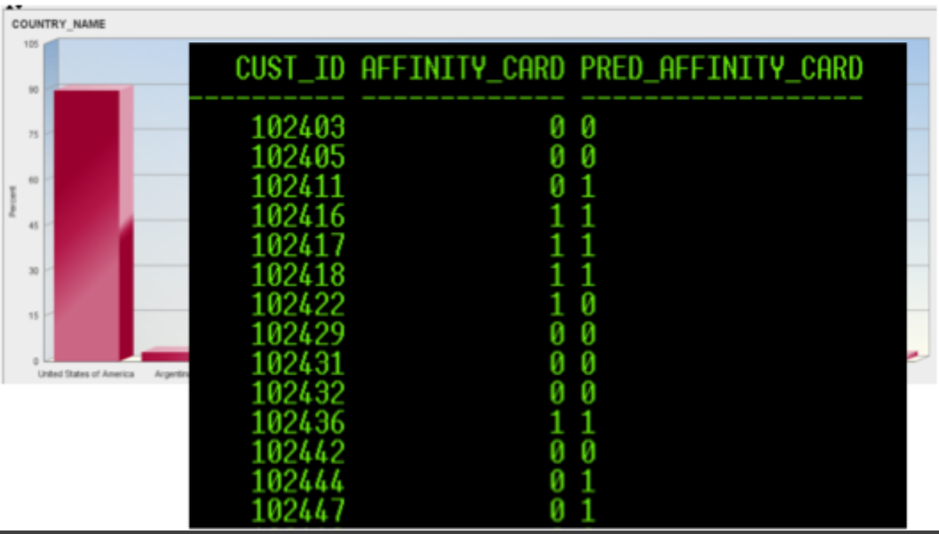
CUST_ID	AFFINITY_CARD	PRED_AFFINITY_CARD	PRED_PROB
102767	1	1	.965276
102770	0	0	.979896
102774	0	0	.784535
102775	0	0	.933800
102779	0	1	.951948
102781	0	1	.753198
102787	0	0	.957232
102791	0	0	.854444
102792	0	0	.643867
102797	0	0	.962632

```
select cust_id, affinity_card,
       PREDICTION( FOR to_char(affinity_card) USING *) OVER ()   pred_affinity_card
from mining_data_build_v;
```



With PQs we can dynamically create new DM models based on an Attribute(s)


```
select cust_id, affinity_card,
       PREDICTION( FOR to_char(affinity_card) USING *) OVER
                                        (PARTITION BY "COUNTRY_NAME") pred_affinity_card
from mining_data_build_v;
```




```
CUST_ID AFFINITY_CARD PRED_AFFINITY_CARD
------- ------------- ------------------
 102403             0 0
 102405             0 0
 102411             0 1
 102416             1 1
 102417             1 1
 102418             1 1
 102422             1 0
 102429             0 0
 102431             0 0
 102432             0 0
 102436             1 1
 102442             0 0
 102444             0 1
 102447             0 1
```

With PQs we can dynamically create new DM models based on an Attribute(s)

A new DM Model will be created for each Country (19)

What about Regression type problems?

```
select customer_id,
       ltv,
       PREDICTION( FOR ltv USING *) OVER ( ) pred_ltv
from   insur_cust_ltv_sample;
```

Use PQs to predict the LTV of customers

CUSTOMER_ID	LTV	PRED_LTV
CU13388	17621.00	17,841
CU13386	22183.00	22,389
CU6607	18805.25	18,732
CU7331	22574.75	22,815
CU2624	17217.25	17,011
CU6389	16140.75	16,266
CU100	24891.25	24,700
CU8653	19238.25	19,002
CU2639	20597.75	20,351
CU1330	19553.00	19,763

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

What about Regression type problems?

```
select customer_id,
       ltv,
       PREDICTION( FOR ltv USING *) OVER ( ) pred_ltv
       (ltv-PREDICTION( FOR ltv USING *) OVER ( )) ltv_diff
from   insur_cust_ltv_sample;
```

Use PQs to predict the LTV of customers

Use PQs to identify and target Customers

CUSTOMER_ID	LTV	PRED_LTV	LTV_DIFF
CU13388	17621.00	17,841	-220
CU13386	22183.00	22,389	-206
CU6607	18805.25	18,732	73
CU7331	22574.75	22,815	-240
CU2624	17217.25	17,011	206
CU6389	16140.75	16,266	-125
CU100	24891.25	24,700	192
CU8653	19238.25	19,002	236
CU2639	20597.75	20,351	247
CU1330	19553.00	19,763	-210

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

What about Regression type problems?

Use PQs to predict the LTV of customers, based on Location

```
select  customer_id,
        ltv,
        PREDICTION( FOR ltv USING *) OVER (PARTITION BY STATE) pred_ltv
        (ltv-PREDICTION( FOR ltv USING *) OVER (PARTITION BY STATE)) ltv_diff
from    insur_cust_ltv_sample;
```

CUSTOMER_ID	LTV	PRED_LTV	LTV_DIFF
CU100	24891.25	24,639	252
CU10006	23638.50	22,899	739
CU10011	35600.50	35,378	223
CU10012	26070.00	23,770	2,300
CU10020	25092.75	24,869	224
CU10025	27149.00	26,944	205
CU10041	27342.50	27,153	189
CU10044	23786.00	24,254	-468
CU1005	25530.25	25,286	244
CU10110	20978.50	20,759	220

www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

What about Regression type problems?

Use PQs to predict the LTV of customers, based on Location & Sex

```
select  customer_id,
        ltv,
        PREDICTION( FOR ltv USING *) OVER (PARTITION BY STATE, SEX) pred_ltv
        (ltv-PREDICTION( FOR ltv USING *) OVER (PARTITION BY STATE, SEX)) ltv_diff
from    insur_cust_ltv_sample;
```

We get slight different Predicted value for LTV.

You would expect this as you are using more detail to build the PQs

	LTV	PRED_LTV	LTV_DIFF
	24891.25	24,666	226
	23638.50	22,728	910
	35600.50	35,351	250
	26070.00	23,770	2,300
	25092.75	24,871	222
CU10025	27149.00	26,902	247
CU10041	27342.50	27,159	183
CU10044	23786.00	24,667	-881
CU1005	25530.25	24,643	887
CU10110	20978.50	20,742	236

#States x #Sex

=19 x 2 = 38 PQ models

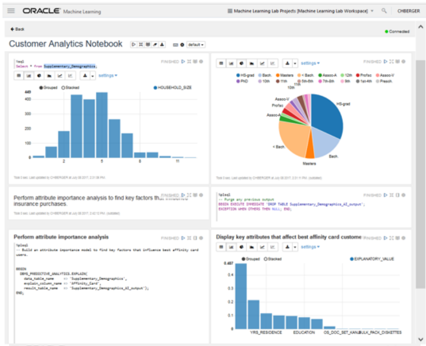
No Extra coding

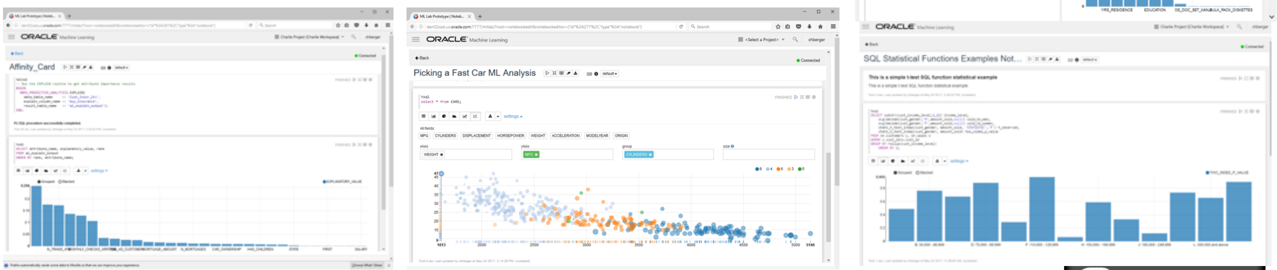
All Dynamic

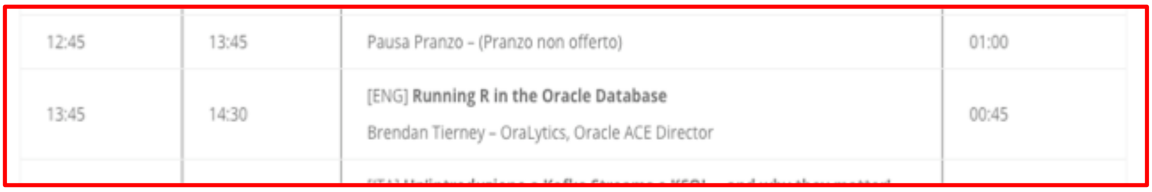
www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com ORALYTICS

Oracle Machine Learning

- Collaborative Notebooks for data scientists
  - Packaged with Autonomous Data Warehouse Cloud (V1)
  - Easy access to shared notebooks, templates, permissions, scheduler, etc.
  - SQL ML algorithms API (V1)
  - Supports deployment of ML analytics







12:45	13:45	Pausa Pranzo – (Pranzo non offerto)	01:00
13:45	14:30	[ENG] **Running R in the Oracle Database** Brendan Tierney – OraLytics, Oracle ACE Director	00:45

ORACLE Database Cloud
High Performance Package

ORACLE Database Cloud
Extreme Performance Package

ORACLE Big Data Cloud Service

ORACLE Data Warehouse Cloud

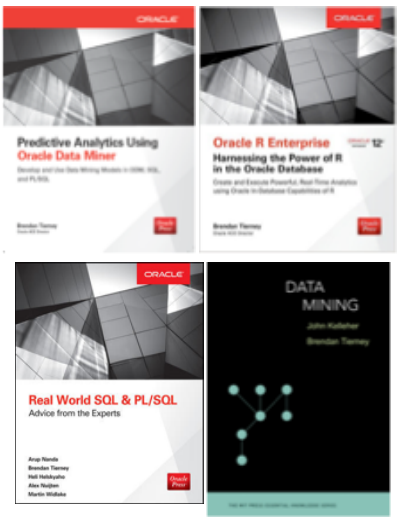
Google brendan tierney data mining

brendan.tierney@oralytics.com

@brendantierney

www.oralytics.com

ie.linkedin.com/in/brendantierney



Word Cloud of the Oracle Advanced Analytics web-pages

http://www.oralytics.com/2015/01/creating-word-cloud-of-oracle-oaa.html

500+ Technical Experts Helping Peers Globally

ORACLE ACE PROGRAM

ORACLE ACE Director

ORACLE ACE

ORACLE ACE Associate

3 Membership Tiers

- Oracle ACE Director
- Oracle ACE
- Oracle ACE Associate

bit.ly/OracleACEProgram

Connect:

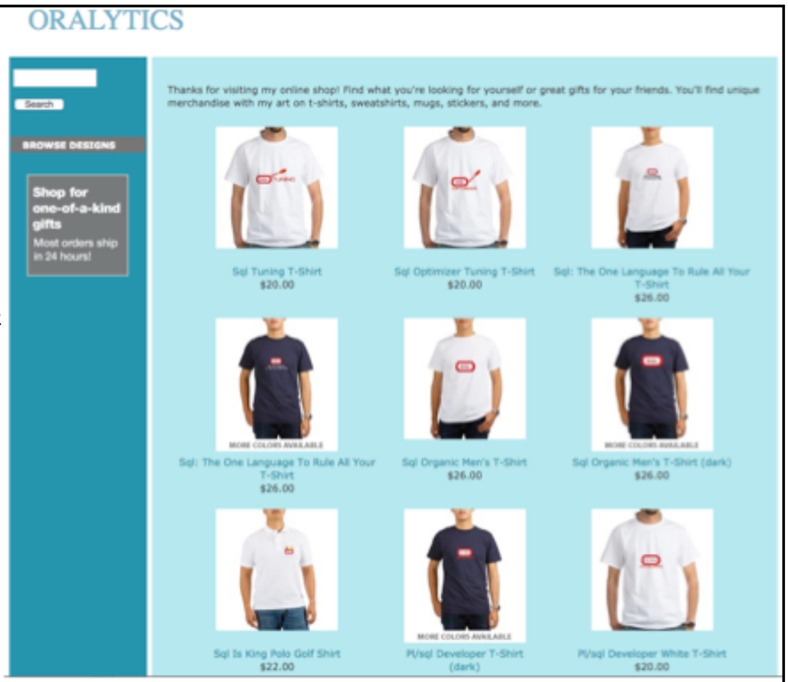
oracle-ace_ww@oracle.com

Facebook.com/oracleaces

@oracleace

ORACLE Developer Community

Nominate yourself or someone you know: acenomination.oracle.com http://www.cafepress.com/oralytics



www.oralytics.com t : @brendantierney e : brendan.tierney@oralytics.com