Introduction to AutoML and Data Science using the Oracle Autonomous Database

Sandesh Rao
VP AIOps for the Autonomous Database

@sandeshr
https://www.linkedin.com/in/raosandesh/
https://www.slideshare.net/SandeshRao4

Safe harbor statement

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, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Statements in this presentation relating to Oracle's future plans, expectations, beliefs, intentions and prospects are "forward-looking statements" and are subject to material risks and uncertainties. A detailed discussion of these factors and other risks that affect our business is contained in Oracle's Securities and Exchange Commission (SEC) filings, including our most recent reports on Form 10-K and Form 10-Q under the heading "Risk Factors." These filings are available on the SEC's website or on Oracle's website at http://www.oracle.com/investor. All information in this presentation is current as of September 2019 and Oracle undertakes no duty to update any statement in light of new information or future events.

Agenda

1. Overview of ML and the Autonomous Database
2. Journey of the DBA to Data Scientist
3. OML Examples
4. AutoML and what’s coming
5. Questions

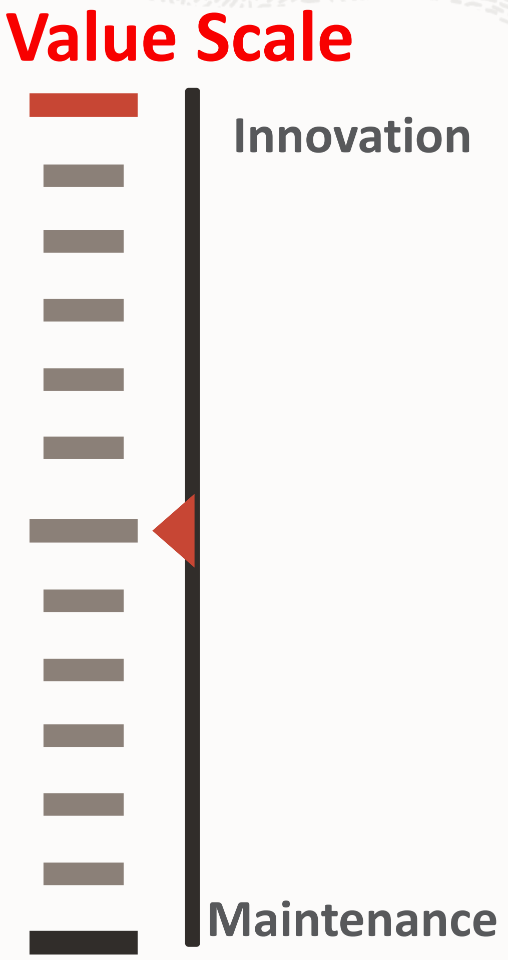
Traditionally DBAs are Responsible for:

Tasks Specific to Business and Innovation

- Architecture, planning, data modeling
- Data security and lifecycle management
- Application related tuning
- End-to-End service level management

Maintenance Tasks

- Configuration and tuning of systems, network, storage
- Database provisioning, patching
- Database backups, H/A, disaster recovery
- Database optimization



Autonomous Database Removes Generic Tasks

Freedom from Drudgery for DBA: More Time to Innovate and Improve the Business

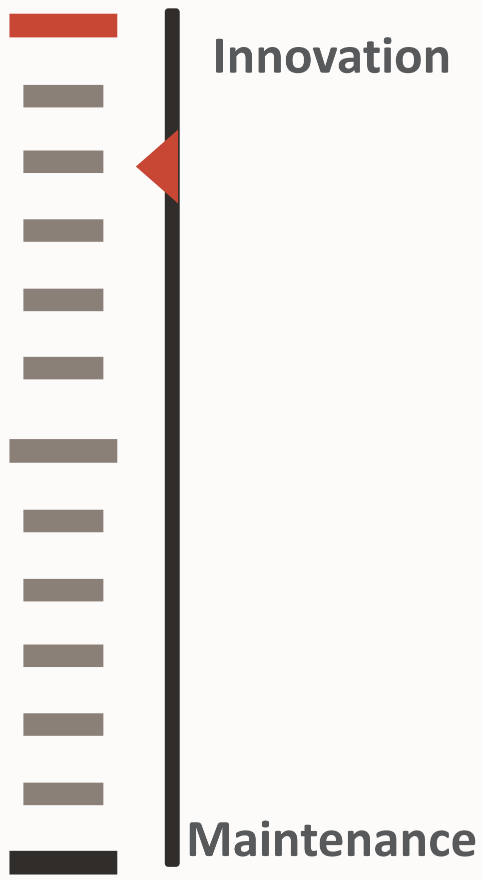
Tasks Specific to Business and Innovation

- Architecture, planning, data modeling
- Data security and lifecycle management
- Application related tuning
- End-to-End service level management

~~Maintenance Tasks~~

- ~~Configuration and tuning of systems, network, storage~~
- ~~Database provisioning, patching~~
- ~~Database backups, H/A, disaster recovery~~
- ~~Database optimization~~

Value Scale



The Evolution of the DBA/Database Developer Role

Data Engineer

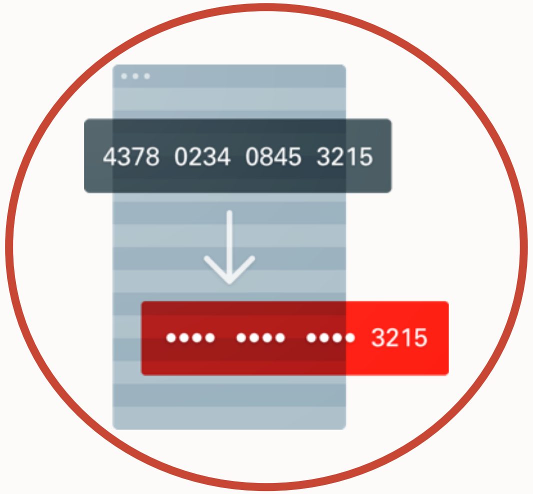
Architecture,
“data wrangler”

Data Security

Data classification,
Data life-cycle mgmt

Machine Learning

Solving data-driven problems
Discovering insights
Making predictions



Application Tuning

SQL tuning,
connection mgmt

Database Developer to Data Scientist Journey

You Are Probably Already Doing Most of This Work!

Data extraction

Data wrangling

Deriving new attributes ("feature engineering")

…

…

…

Import predictions & insights

Translate and deploy ML models

Automate

Typically 80% of the work

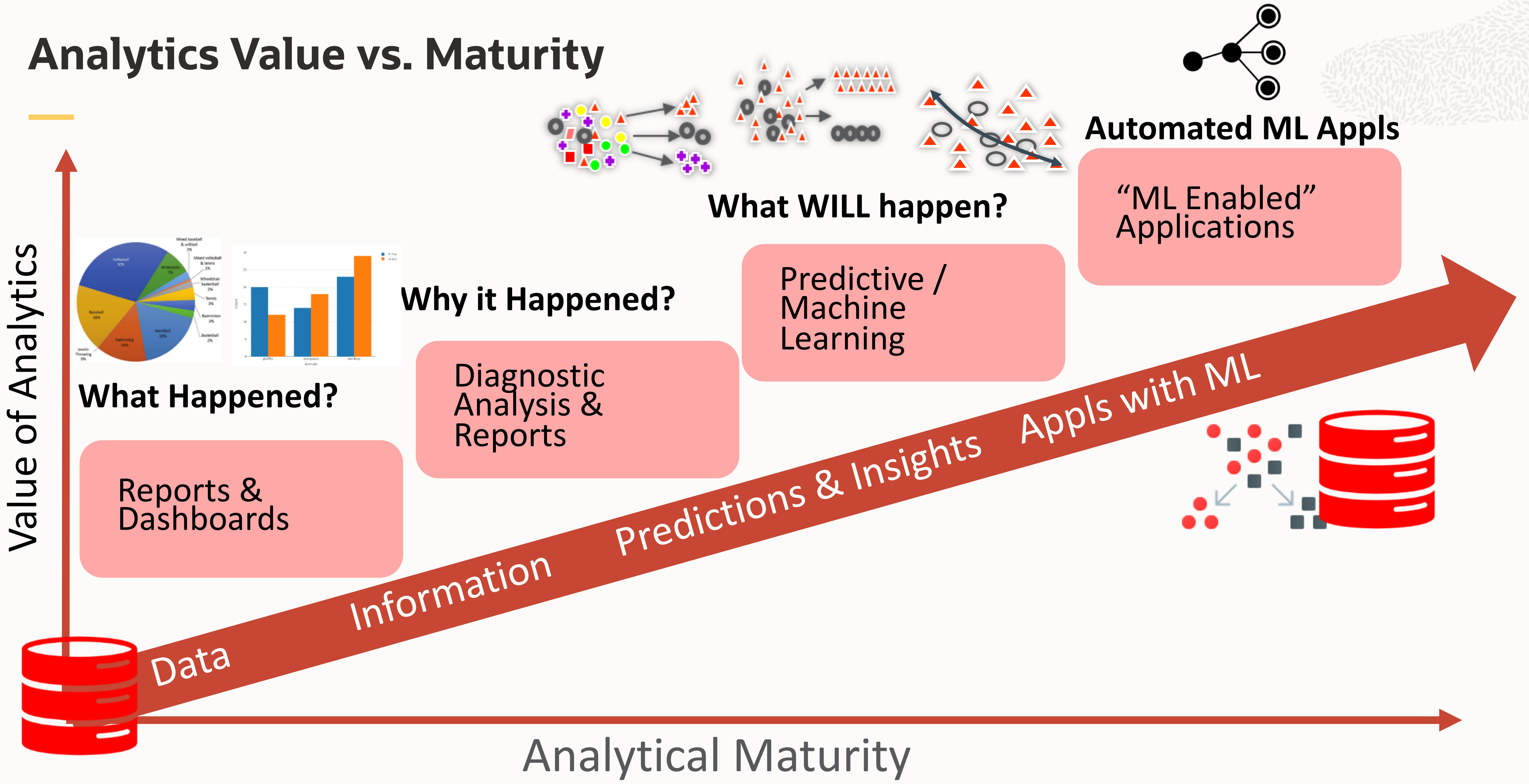
Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy[1]

Eliminated or minimized with Oracle

Data Management platform becomes combine/hybrid DM + machine learning platform

[1] - https://www.infoworld.com/article/3228245/data-science/the-80-20-data-science-dilemma.html

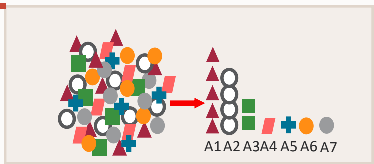
Analytics Value vs. Maturity



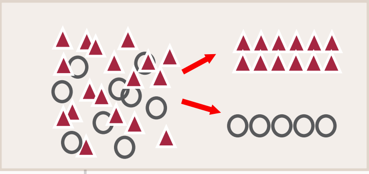
What is Machine Learning?

Algorithms ***automatically*** sift through large amounts of data to discover hidden patterns, new insights and make predictions

Supervised Learning

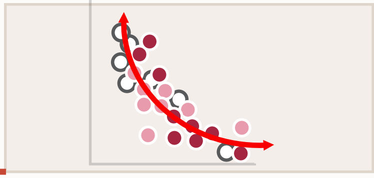


Identify most important factor (Attribute Importance)



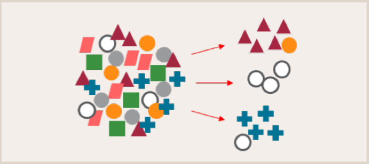
Predict customer behavior (Classification)

Find profiles of targeted people or items (Classification

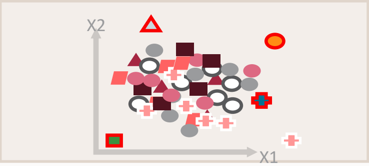


Predict or estimate a value (Regression)

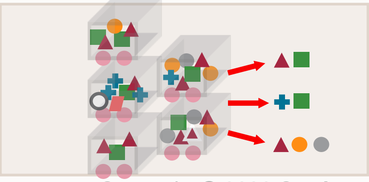
Unsupervised Learning



Segment a population (Clustering)



Find fraudulent or “rare events” (Anomaly Detection)



Determine co-occurring items in a “basket” (Associations)

CRISP-DM Methodology

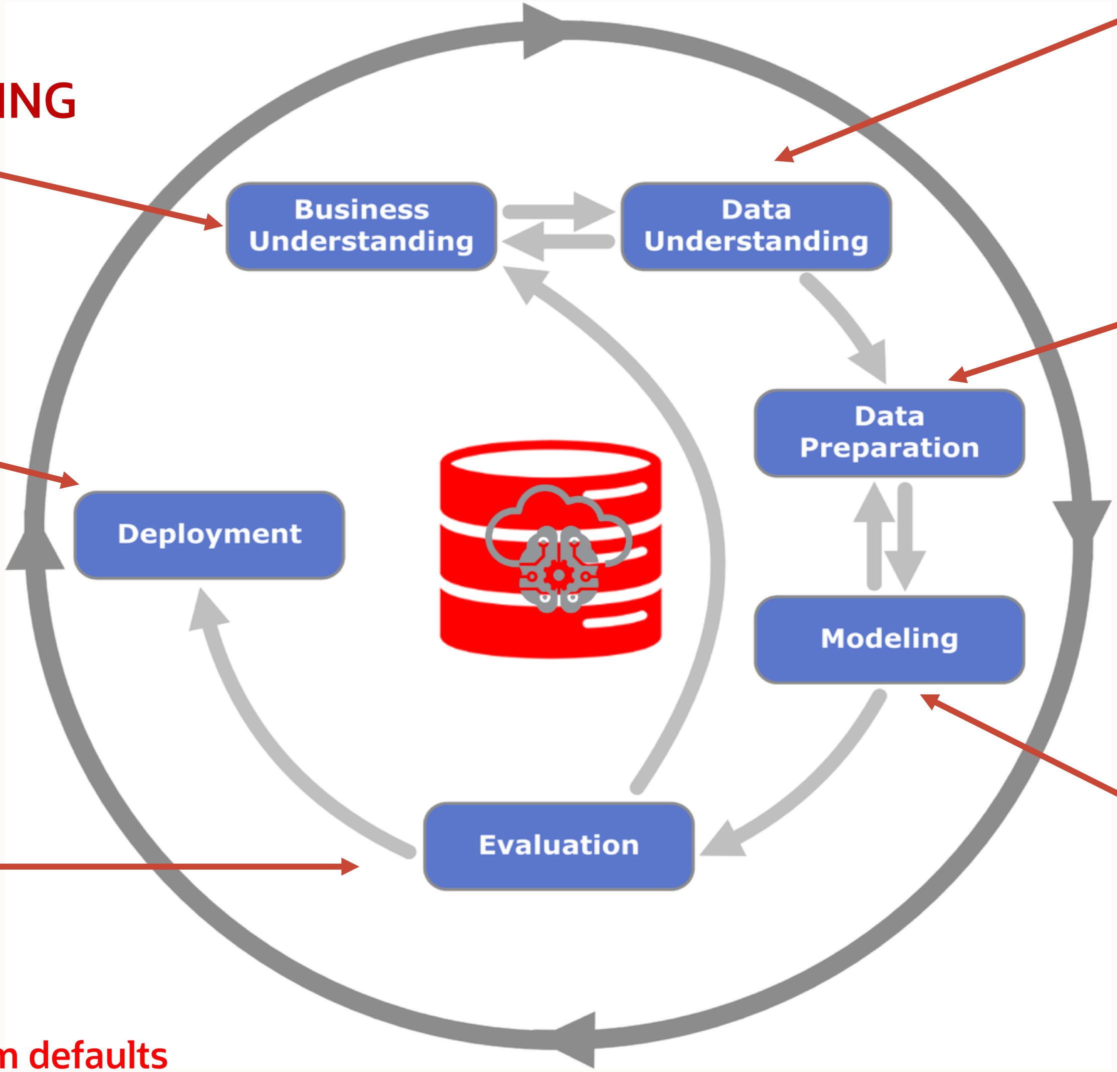
Six Major Steps

BUSINESS UNDERSTANDING
Well-defined business problem

DEPLOYMENT
In-DB ML model apply
- Real-time ML apply
- In-database, REST

Embed methodology
- Applications
- Dashboards

EVALUATION
Model evaluation
Model comparison
Model selection



DATA UNDERSTANDING
Assemble the "right data"
Data profiling
- Data visualization
- Univariate statistics/group by
- Bi-variate statistics

DATA PREPARATION
Sampling/Stratified
Algorithm req'd transforms
- Auto Data Preparation
- Missing Values, Binning, Normalization, etc.
- Unstructured data
- Aggregations

Domain specific transforms
- "Engineered Features"

Features Selection

MODELING
Algorithm settings/defaults
- Stratified sampling
- Feature selection
- Build model(s)

* Automated and/or system defaults

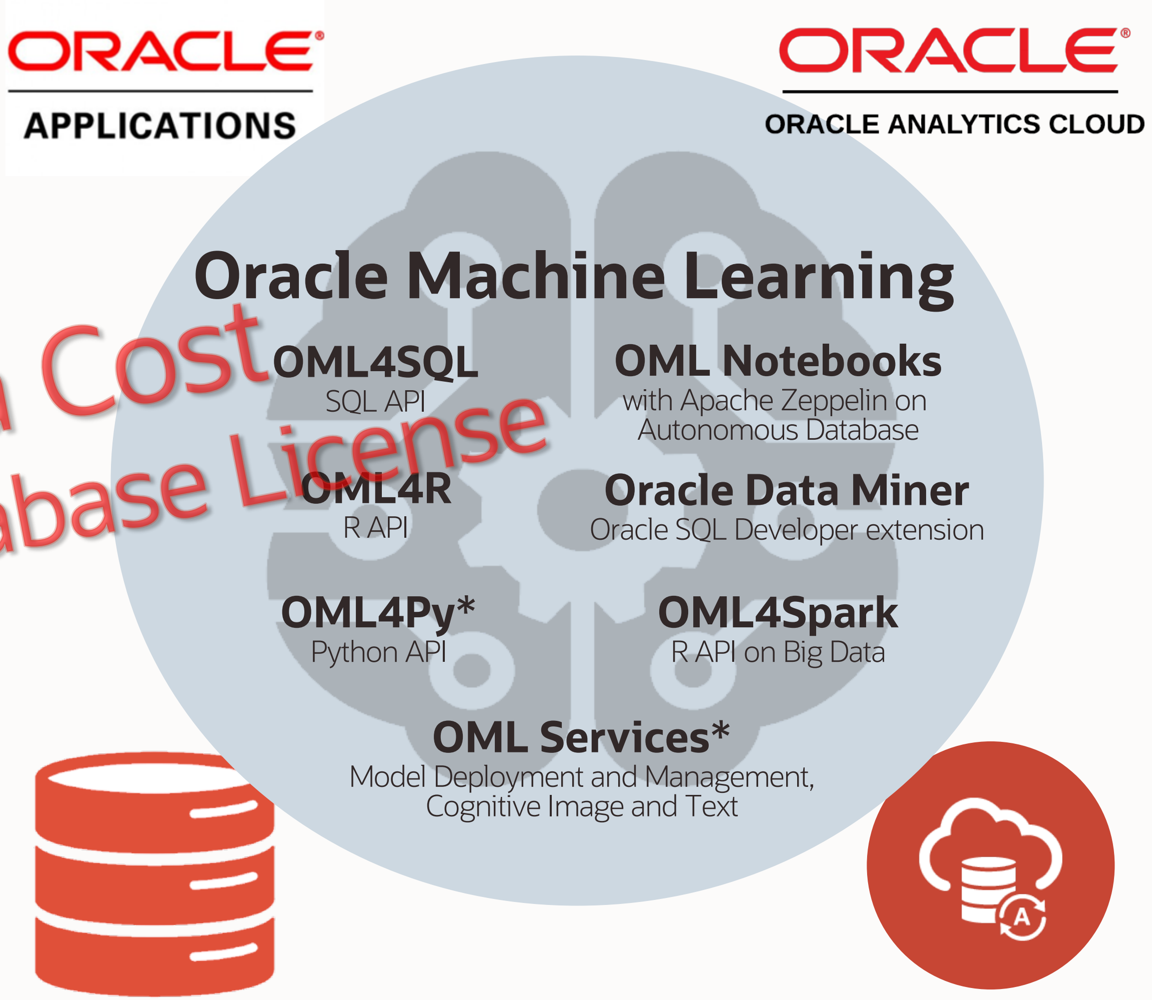
 https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Oracle Machine Learning

Oracle Machine Learning extends Oracle Database(s) and enables users to build “AI” applications and analytics dashboards

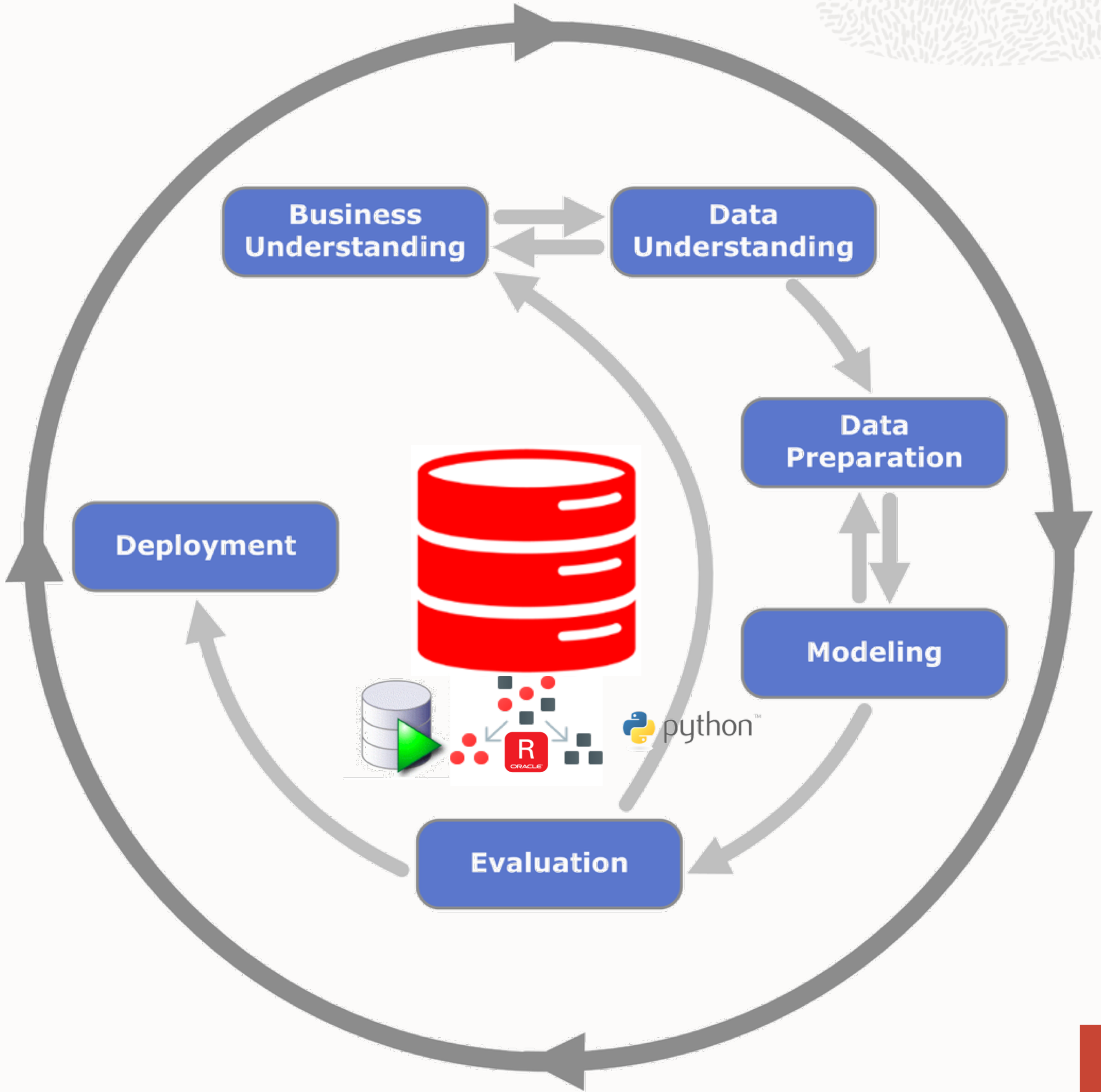
OML delivers powerful in-database machine learning algorithms, automated ML functionality via SQL APIs and integration with open source Python* and R.

No Extra Cost
Included in Database License



* Coming soon

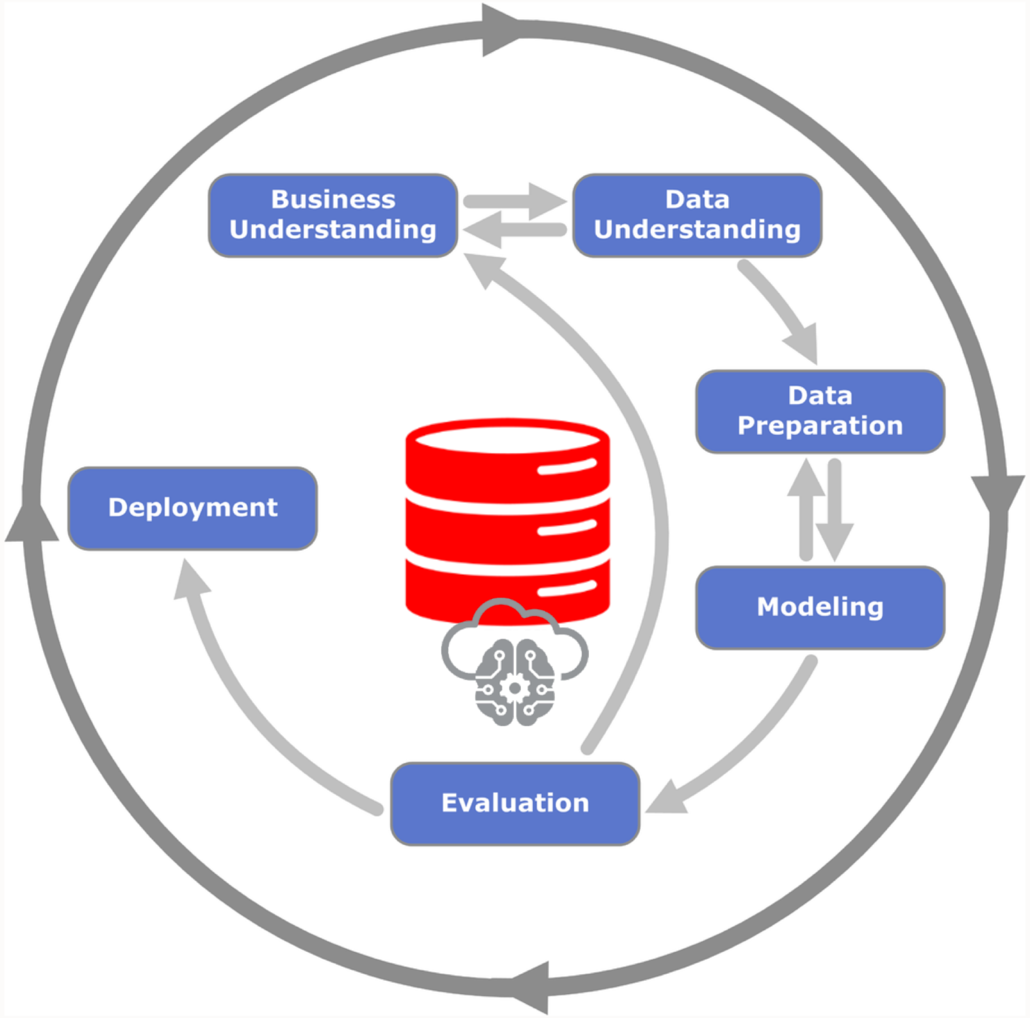
Database Developer to Data Scientist Journey



Database Developer to Data Scientist Journey

Six Major Steps (Oracle Machine Learning POV)

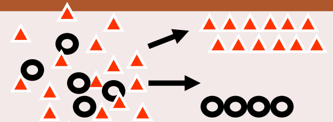
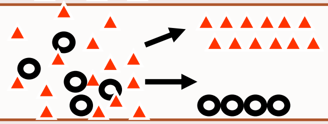
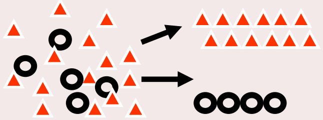
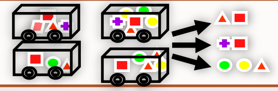
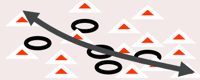
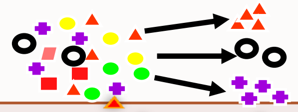
- Business Understanding—Week 1
- Data Understanding—Week 2
- Data Preparation—Week 3
- Modeling (ML)—Week 4
- Evaluation—Week 5
- Deployment—Week 6



https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Week 1—Business Understanding

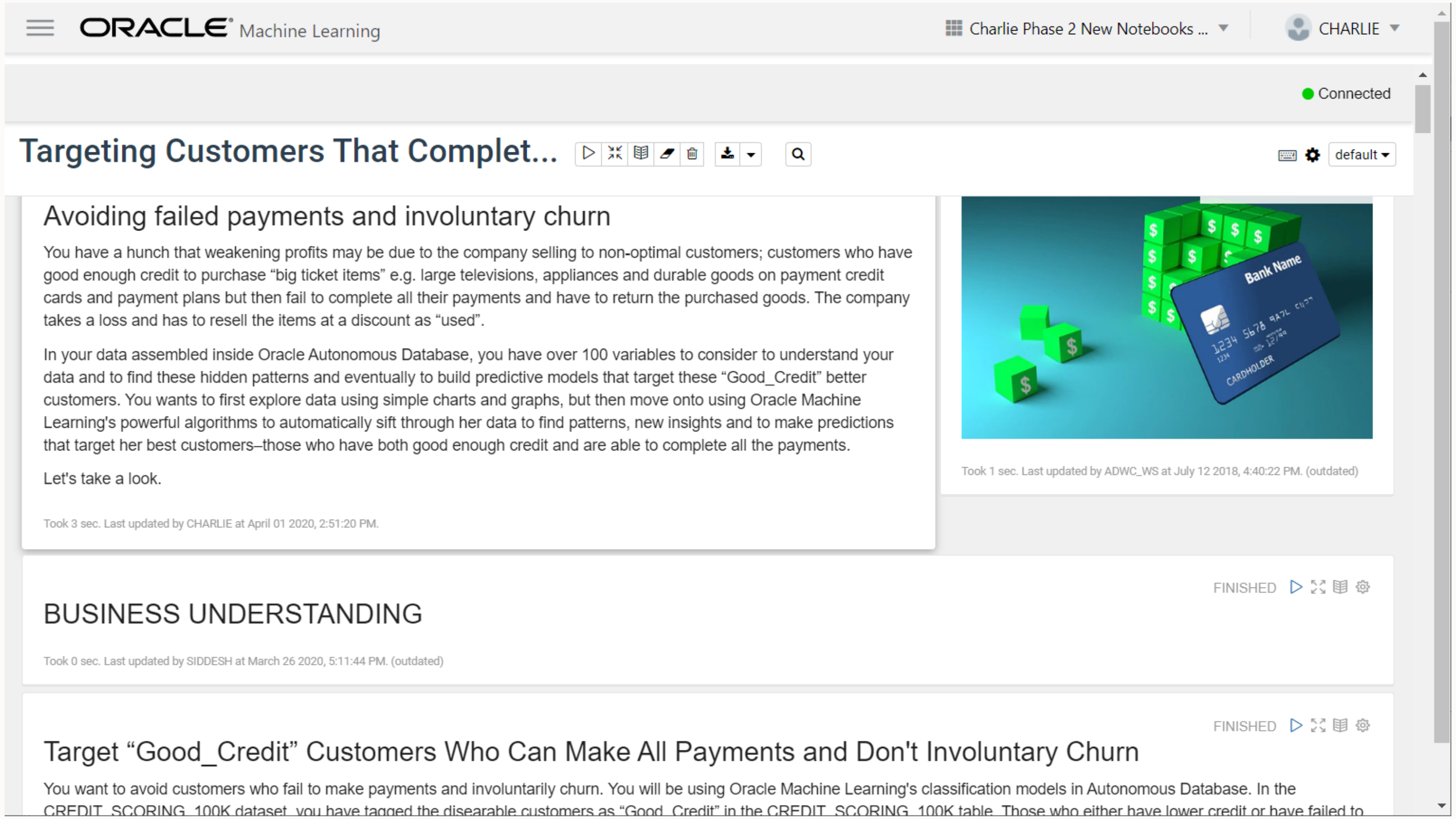
Start with a Well-Defined Business Problem Statement

Poorly Defined	Better	Data Mining Technique
Predict employees that leave	• Based on past employees that voluntarily left: • Create New Attribute EmplTurnover → 0/1	
Predict customers that churn	• Based on past customers that have churned: • Create New Attribute Churn → YES/NO	
Target "best" customers	• Recency, Frequency Monetary (RFM) Analysis • Specific Dollar Amount over Time Window: • Who has spent $500+ in most recent 18 months	
How can I make more $$?	• What helps me sell soft drinks & coffee?	
Which customers are likely to buy?	• How much is each customer likely to spend?	
Who are my "best customers"?	• What descriptive "rules" describe "best customers"?	
How can I combat fraud?	• Which transactions are the most anomalous? • Then roll-up to physician, claimant, employee…	

Week 1—Business Understanding

Be Extremely Specific in your Problem Statement

Target “best” customers who have **GOOD CREDIT and make payments**



Week 2—Data Understanding

Review the Data; Does it Makes Sense?

Credit Score Predictions

Took 4 sec. Last updated by CHARLIE at February 07 2019, 3:29:43 PM.

STEP 5: Review Credit Scoring Data

```
%sql

/* This shows the credit scoring data we will use historical data to predict the likelihood o

Select * from credit_scoring_100k_v where rownum < 100
```

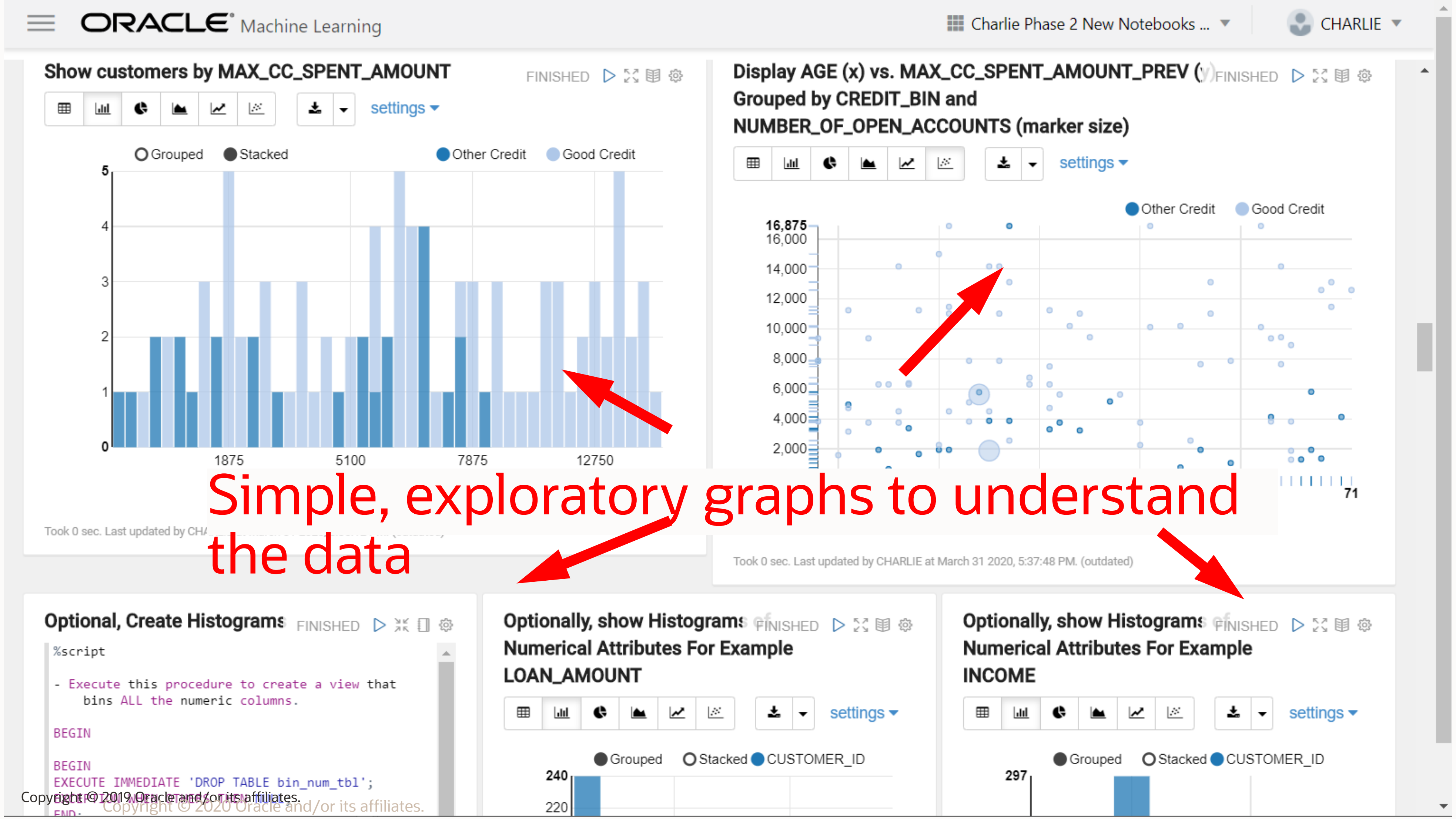
Are AGEs all positive, 0-120?
Are INCOME values weekly or monthly?
Are the LOAN_AMOUNTS reasonable?
Etc….

CUSTOMER_ID	AGE	INCOME	MARITAL_STATUS	NUMBER_OF_LIABLES	WEALTH	EDUCATION_LEVEL	TENURE	LOAN_TYPE	LOAN_AMOUNT	LOAN_LENGTH	GENDER	RE
41908	27	3750	Single	5	Average	Master's Degree	15	Education	15000	12	Male	Ae
54255	24	2150	Married	2	Average	Master's Degree	5	Housing	55000	72	Male	Bl
46270	37	2150	Married	2	Average	Master's Degree	24	Education	20000	12	Female	Me
54244	47	2150	Divorced	1	Average	Master's Degree	17	Need	35000	12	Male	Ce
49074	21	2750	Married	2	Average	Master's Degree	24	Housing	95000	108	Female	Ma
10647	63	2750	Married	5	Poor	Bachelor's Degree	7	Need	30000	24	Female	Ce
9535	28	2750	Married	3	Poor	Bachelor's Degree	11	Housing	80000	60	Male	Sc
34680	67	2150	Married	3	Poor	Bachelor's Degree	11	Need	15000	12	Male	Ma

Took 5 sec. Last updated by CHARLIE at February 07 2019, 3:29:56 PM.

Week 2—Data Understanding

Review the Data; Does it Makes Sense?

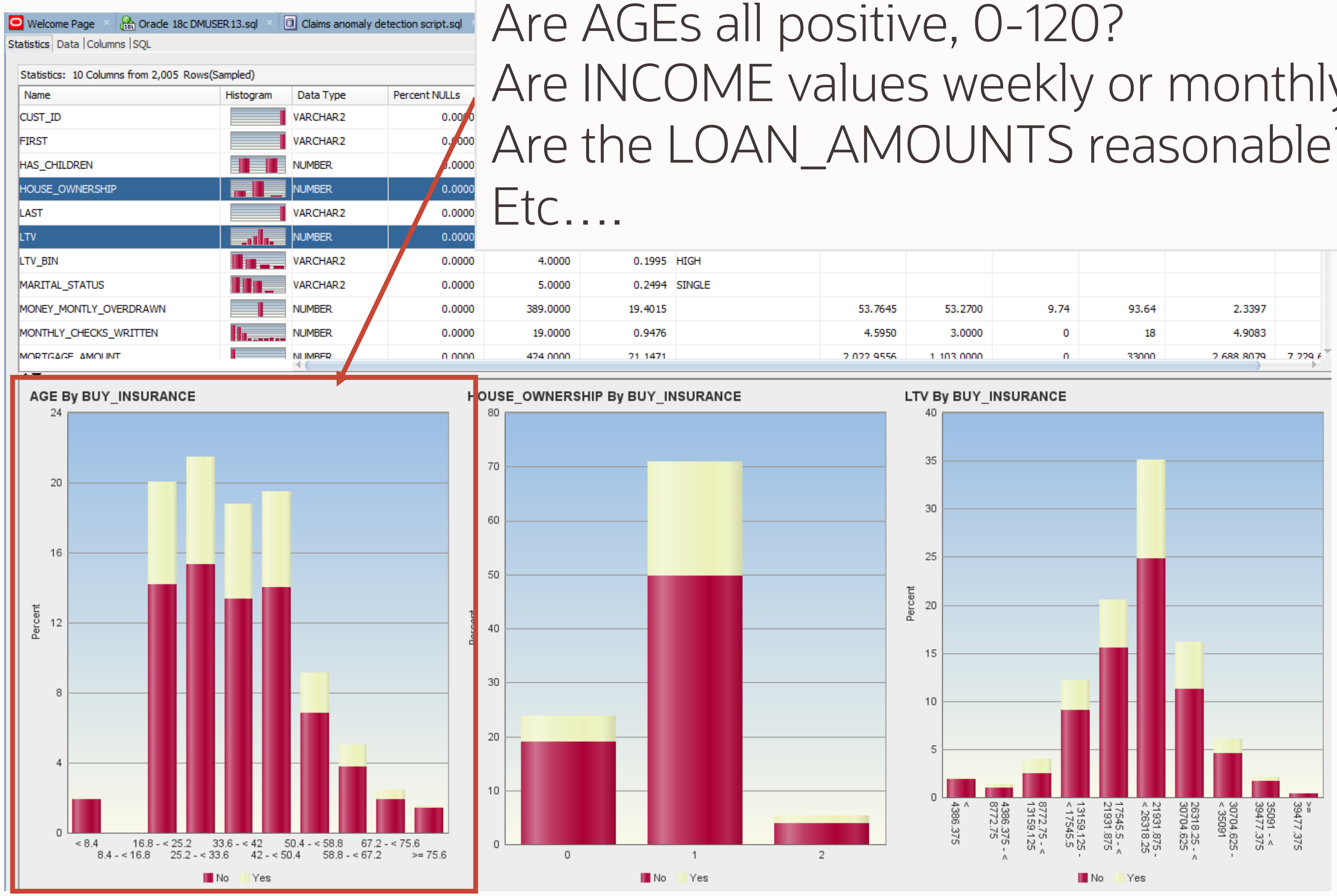


Simple, exploratory graphs to understand the data

Week 2—Data Understanding

Review the Data; Does it Makes Sense?

Are AGEs all positive, 0-120?
Are INCOME values weekly or monthly?
Are the LOAN_AMOUNTS reasonable?
Etc….



Week 3—Data Preparation

Prepare the Data, Create New Derived Attributes or “Engineered Features”

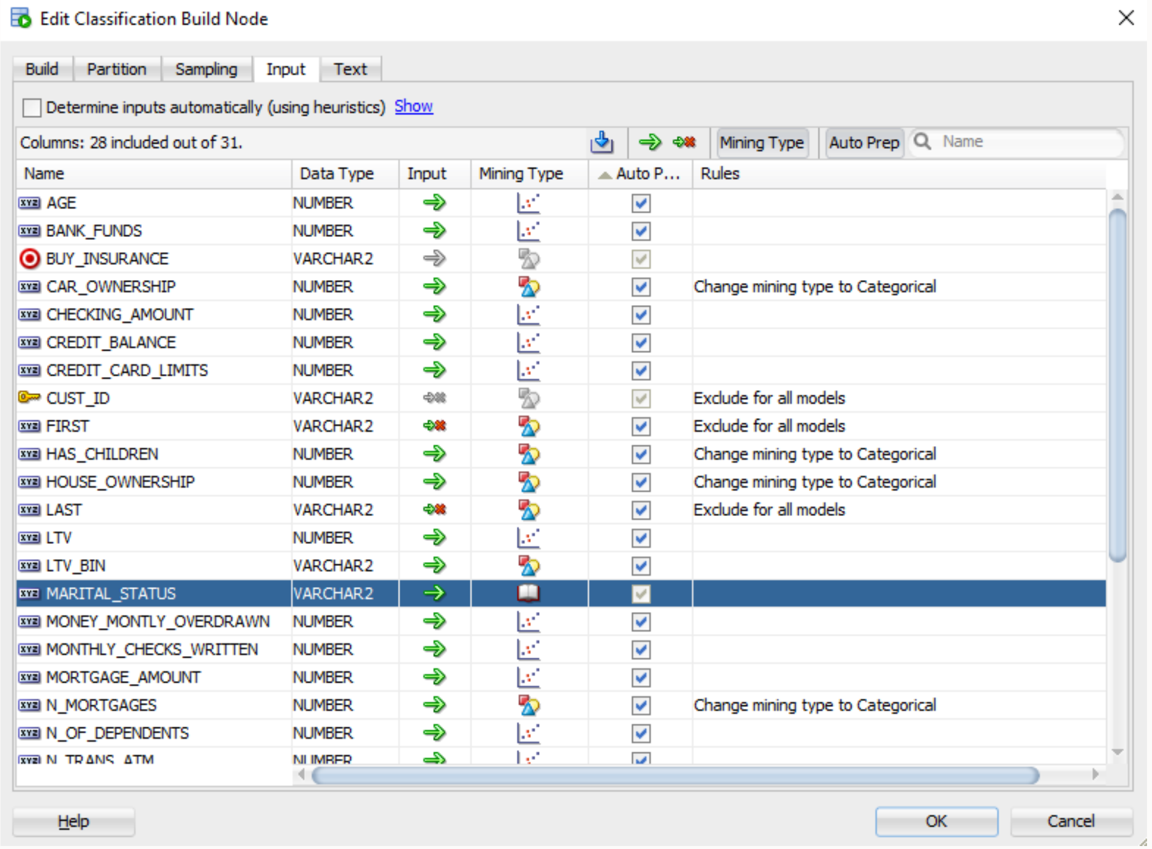
Source Attribute		New Attribute/”Engineered Feature”
Date of Birth	→	AGE
Address	→	DISTANCE_TO_DESTINATION
	→	COMMUTE_TIME
Call detail records (CDRs)	→	#_DROPPED_CALLS
	→	PERCENT_INTERNATIONAL
Salary	→	PERCENT_VS_PEERS
Purchases	→	TOTALS_PER_CATEGORY *(e.g. Food, Clothing)*

Week 3—Data Preparation

Prepare the Data, Create New Derived Attributes or “Engineered Features”

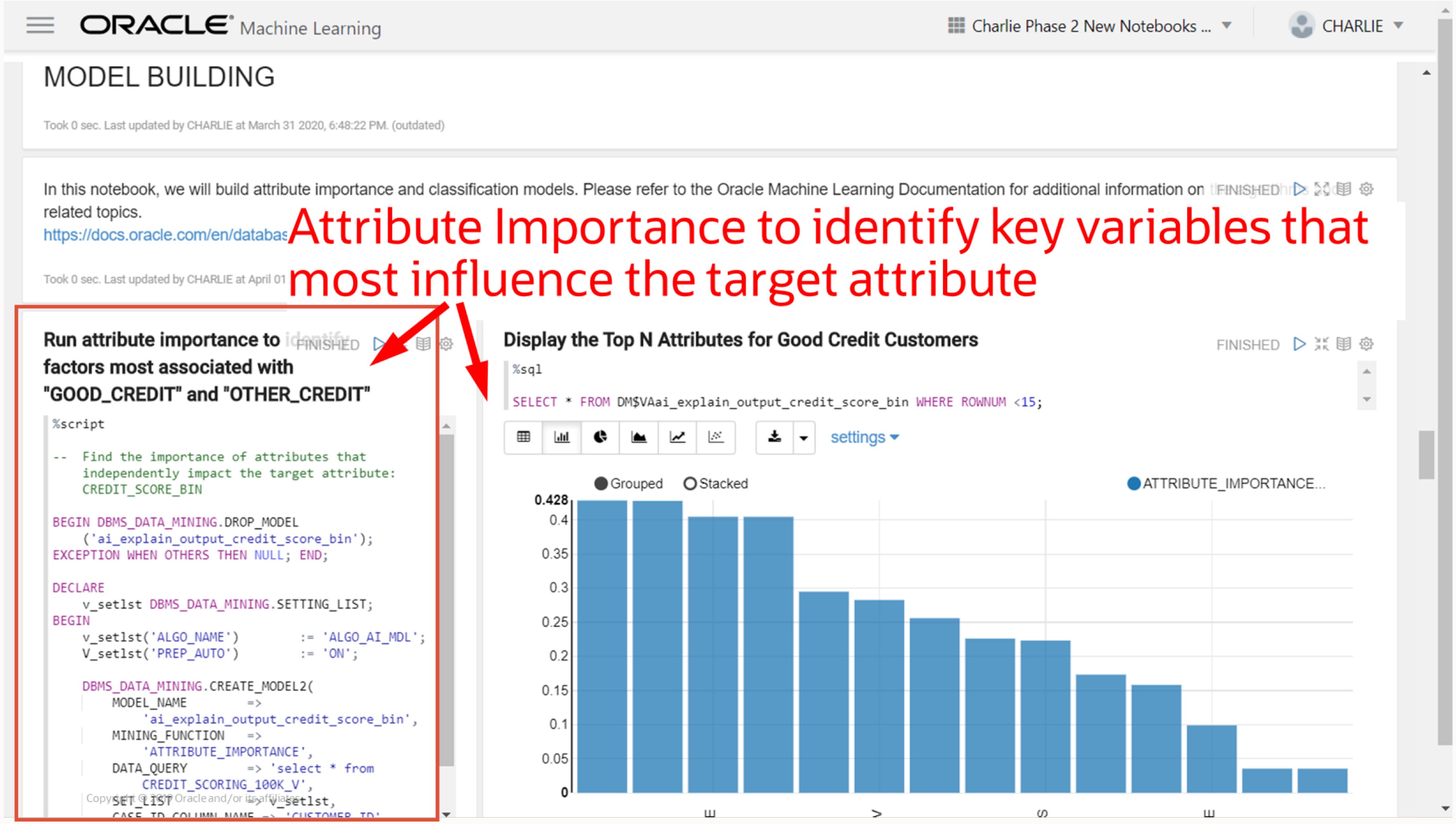
Oracle Machine Learning’s Auto Data Prep (ADP) and ML algorithms are designed with intelligent defaults and can automatically deal with:

- Missing values
- Outliers
- Binning
- Too many distinct values
- Too many constants
- Trans data/aggregations
- Unstructured data
- Correlated data



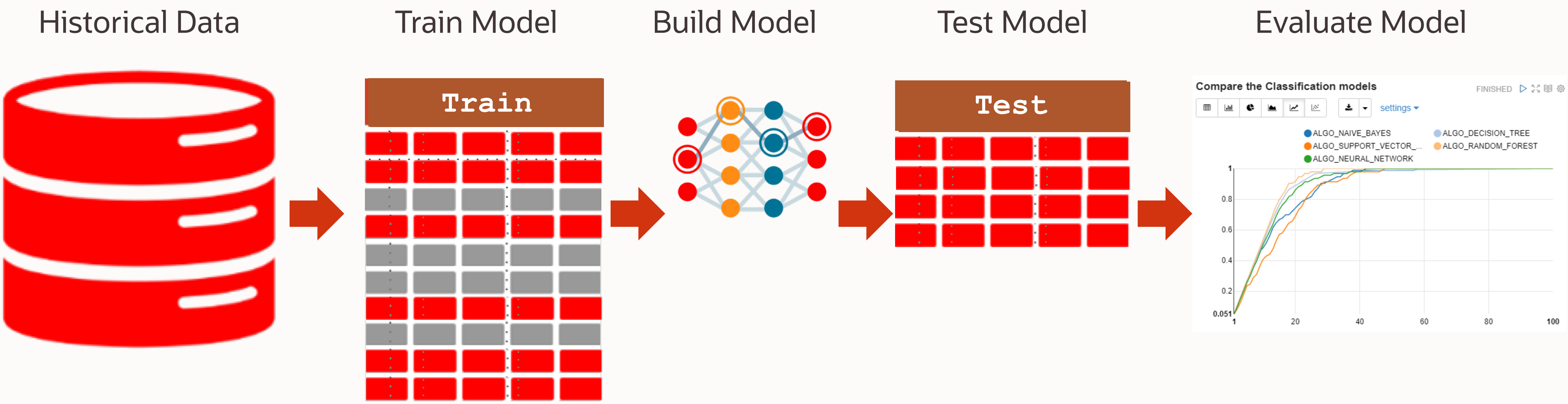
Week 4—Modeling (Machine Learning)

First, Identify the Key Attributes That Most Influence the Target Attribute



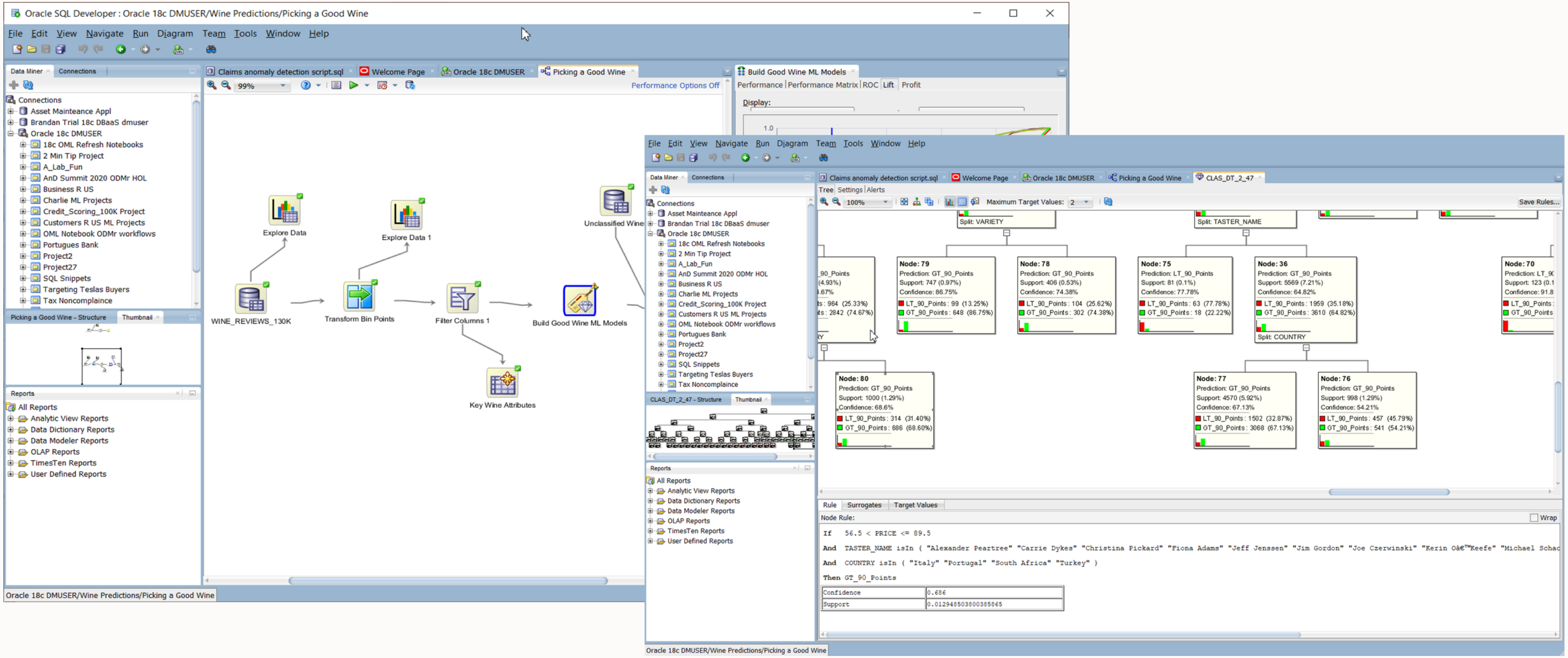
Week 4—Modeling (Machine Learning)

Training and Testing ML Models using 60/40% Random Samples



Week 4—Modeling (Machine Learning)

Build multiple models with different algorithms and settings



Week 5—Model Evaluation (ML)

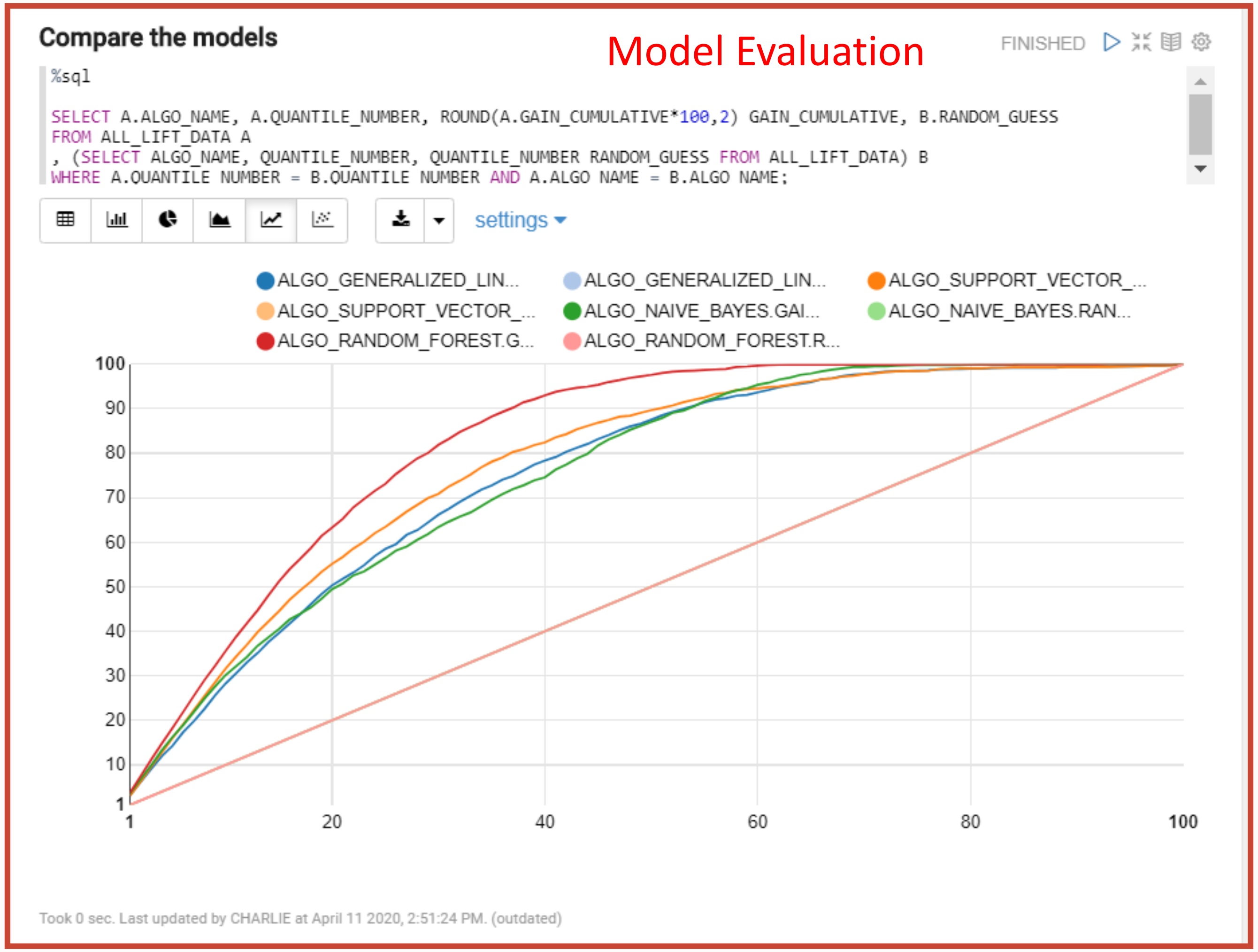
Next, test model accuracy

Randomly selected “hold out” sample of data that was used to train the ML model

Compute Cumulative Gains, Lift, Accuracy, etc.

Review the attributes used in the model and model coefficients

Make sure the model makes sense



Week 6—Deployment

Apply the Models to Predict “Best Customers”

Simple SQL Apply scripts run 100% inside the Database for immediate ML model deployment

Apply the Oracle Machine Learning Model to New Customers to Show Customers Most Likely to Have Good Credit

FINISHED

%sql

Model Apply/”Scoring”

```
select a.customer_id
    , a.prob_Credit_Score_Bin
    , b.age, b.income, b.tenure, b.loan_type, b.loan_amount, b.occupation, b.education_level, b
        .marital_status
from (select * from (select Customer_id, round(prob_Credit_Score_Bin *100,2) prob_Credit_Score_Bin from
    (select Customer_ID, prediction_probability(N1_CLASS_MODEL, NULL using *) prob_Credit_Score_Bin from
    credit_scoring_new_cust_v))) a
, credit_scoring_100k_v b
where a.customer_id = b.customer_id
order by a.prob_Credit_Score_Bin desc
```

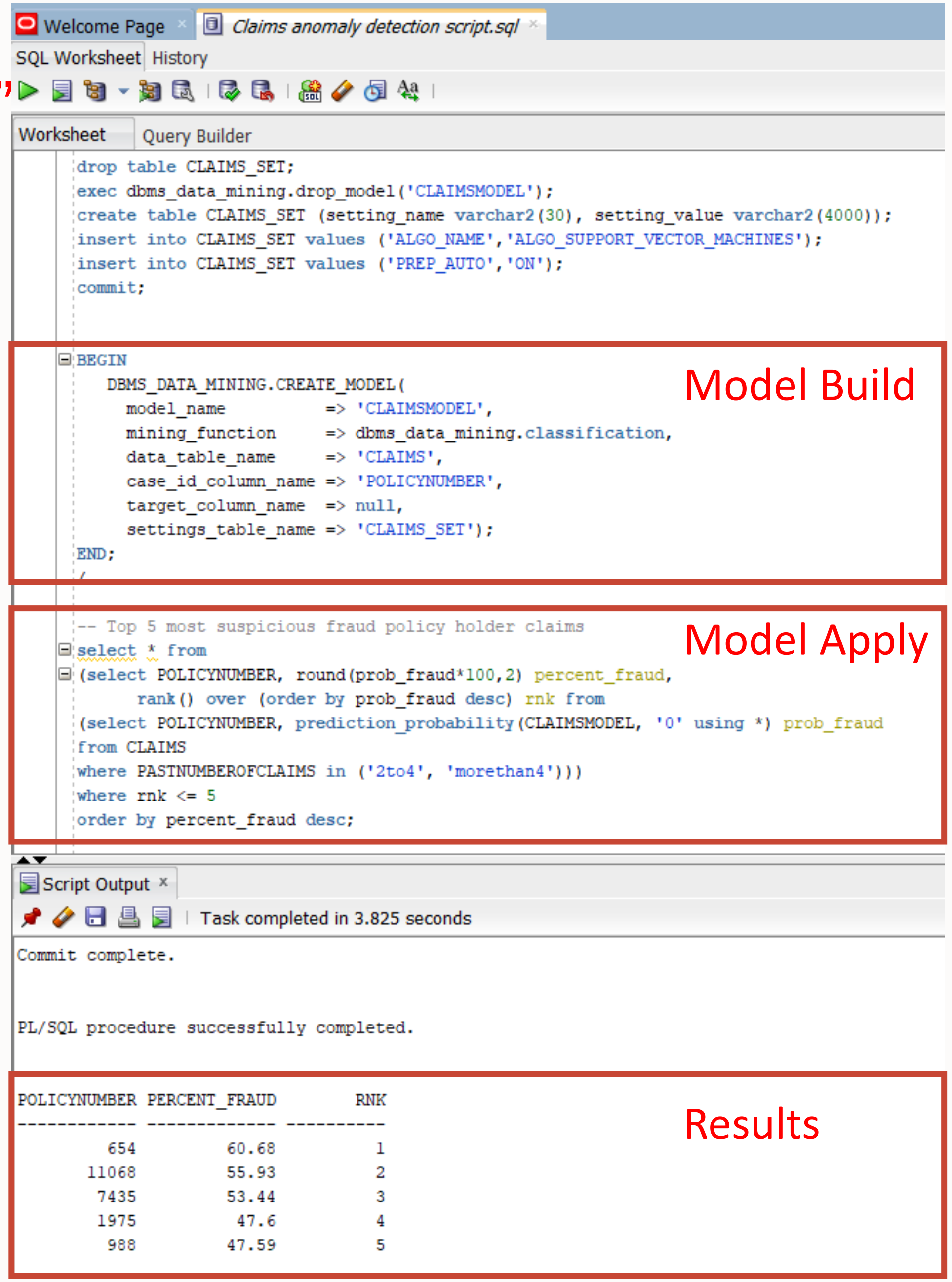
CUSTOMER_ID	PROB_CREDIT_SCORE_BIN	AGE	INCOME	TENURE	LOAN_TYPE
34673	100	31	5250	8	Education
77936	100	37	6250	6	Need
56154	100	45	4250	10	Need
11610	100	28	6250	3	Housing
56733	100	63	4250	4	Housing
57999	100	54	4250	33	Housing

Took 0 sec. Last updated by ADWC_WS2 at October 17 2018, 2:51:03 PM. (outdated)

Week 6—Deployment

Apply the Models to Predict “Best Customers”

Simple SQL Apply scripts run 100% inside the Database for model build, model apply and immediate ML model deployment



Congratulations!
Almost there ☺

OML examples

Statistical Functions

Simple SQL Syntax—Statistical Comparisons (t-tests)

Compare AVE Purchase Amounts Men vs. Women Grouped_By INCOME_LEVEL

```
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_INDEPU(cust_gender, amount_sold, 'STATISTIC', 'F') t_observed,
       STATS_T_TEST_INDEPU(cust_gender, amount_sold) two_sided_p_value
  FROM customers c, sales s
  WHERE c.cust_id = s.cust_id
  GROUP BY ROLLUP(cust_income_level)
  ORDER BY income_level, sold_to_men, sold_to_women, t_observed;
```

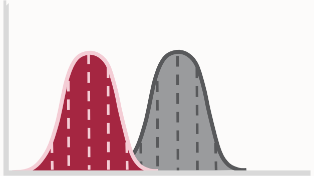
Query Result ×

SQL | All Rows Fetched: 14 in 1.523 seconds

	INCOME_LEVEL	SOLD_TO_MEN	SOLD_TO_WOMEN	T_OBSERVED	TWO_SIDED_P_VALUE
1	A: Below 30,000	105.2834897729...	99.42814466653473...	-2.05425922984...	0.039964704379552678
2	B: 30,000 - 49,999	102.5965095067...	109.8296418272003...	2.969223321889...	0.0029877419365879512
3	C: 50,000 - 69,999	105.6275880730...	110.1279310121247...	2.349685400926...	0.018792276771129993
4	D: 70,000 - 89,999	106.6302994897...	110.4728699326023...	2.268392806338...	0.023307831257217089
5	E: 90,000 - 109,999	103.3967414937...	101.6104162583700...	-1.26035091954...	0.20754566236328209
6	F: 110,000 - 129,999	106.7647596205...	105.9813119482142...	-0.60580010770...	0.54464855287037528
7	G: 130,000 - 149,999	108.8775321810...	107.3137698570293...	-0.85219780969...	0.39410775484348759
8	H: 150,000 - 169,999	110.9872579252...	107.1521911799573...	-1.94514858879...	0.051762623899376248
9	I: 170,000 - 189,999	102.8082379709...	107.4355601412162...	2.149669205899...	0.031587875078399455
10	J: 190,000 - 249,999	108.0405638372...	115.3433560297627...	2.547498669040...	0.010854966021230945
11	K: 250,000 - 299,999	112.3779929260...	108.1960973300511...	-1.41155136806...	0.15809167565415438
12	L: 300,000 and above	120.9702345758...	112.2163421398336...	-2.07261936500...	0.038225611271820475
13	(null)	106.6637691587...	107.2763858209415...	1.078537818864...	0.28079420737509053
14	(null)	107.1218447412...	113.8044098205854...	0.689462437157...	0.4905957646106357

STATS_T_TEST_INDEPU (SQL) Example;

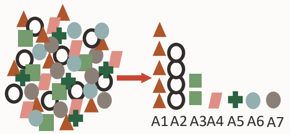
P_Values < 05 show statistically significantly differences in the amounts purchased by men vs. women



OAA Model Build and Real-time SQL Apply Prediction

Simple SQL Syntax—Attribute Importance - ML Model Build (PL/SQL)

```
BEGIN
   DBMS_DATA_MINING.CREATE_MODEL(
     model_name           => 'BUY_INSURANCE_AI',
     mining_function              => DBMS_DATA_MINING.ATTRIBUTE_IMPORTANCE,
     data_table_name              => 'CUST_INSUR_LTV',
     case_id_column_name          => 'cust_id',
     target_column_name           => 'BUY_INSURANCE',
     settings_table_name          => 'Att_Import_Mode_Settings');
END;
/
```



Model Results (SQL query)

```
SELECT attribute_name, rank , attribute_value
FROM BUY_INSURANCE_AI
ORDER BY rank, attribute_name;
```

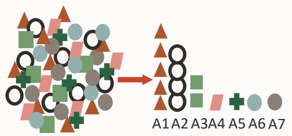
ATTRIBUTE_NAME	RANK	ATTRIBUTE_VALUE
BANK_FUNDS	1	0.2161
MONEY_MONTLY_OVERDRAWN	2	0.1489
N_TRANS_ATM	3	0.1463
N_TRANS_TELLER	4	0.1156
T_AMOUNT_AUTOM_PAYMENTS	5	0.1095

OML for R Model Build

Simple R Language Syntax—Attribute Importance

ML Model Build (R)

```
> ore.odmAI (BUY_INSURANCE ~ ., CUST_INSUR_LTV)

Call:
ore.odmAI(formula = BUY_INSURANCE ~ ., data = CUST_INSUR_LTV)
```



Model Results (R)

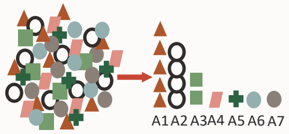
```
Importance:
                         importance     rank
BANK_FUNDS               0.2161187797    1
MONEY_MONTLY_OVERDRAWN   0.1489347141    2
N_TRANS_ATM              0.1463026512    3
N_TRANS_TELLER           0.1155879786    4
T_AMOUNT_AUTOM_PAYMENTS  0.1095178647    5
```

OML for Python Model Build—*Coming soon!*

Simple Python Language Syntax—Attribute Importance

ML Model Build (Python)

```
> ai_mod = ai(**setting)   # Create AI model object
> ai_mod = ai_mod.fit(train_x, train_y)
```



Model Results (Python)

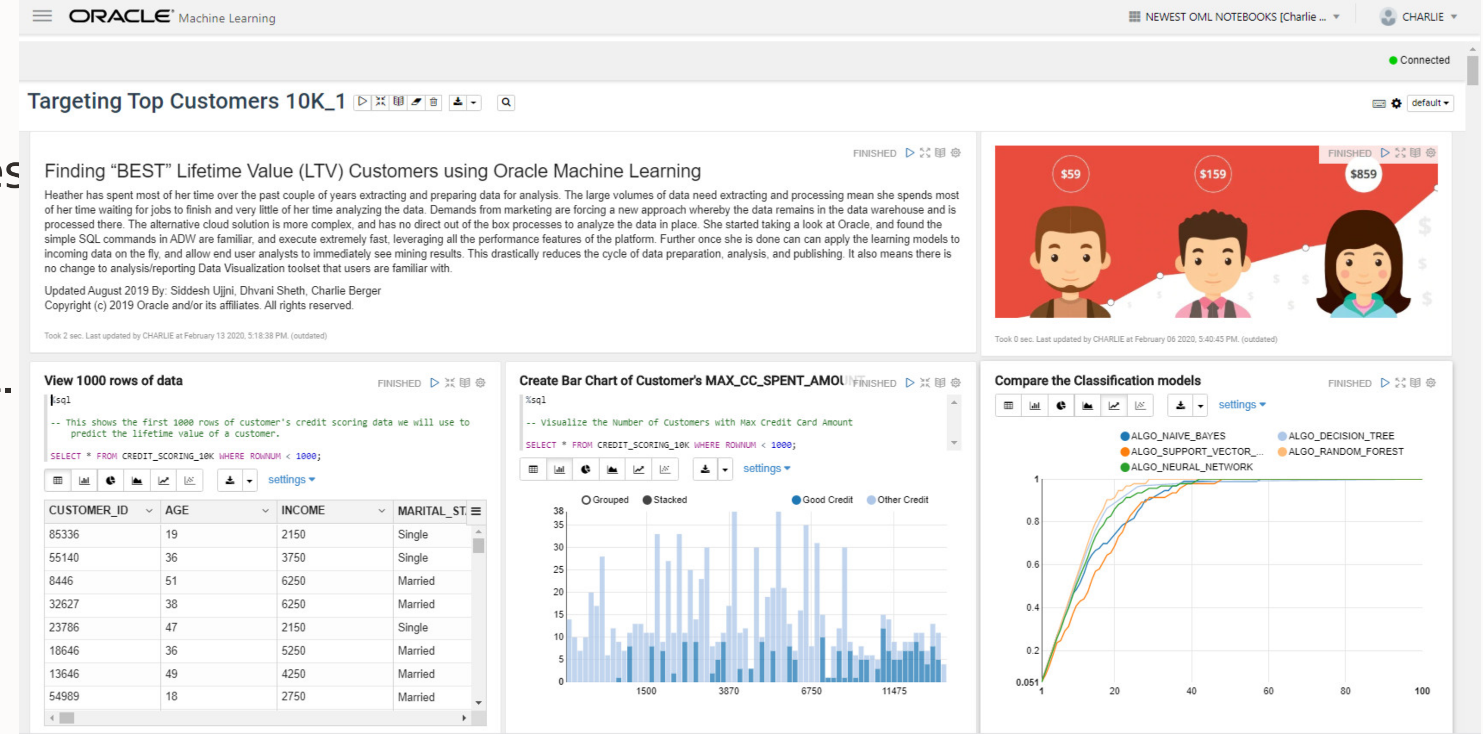
```
Importance:

variable                 importance      rank
BANK_FUNDS                0.2161187797     1
MONEY_MONTLY_OVERDRAWN    0.1489347141     2
N_TRANS_ATM               0.1463026512     3
N_TRANS_TELLER            0.1155879786     4
T_AMOUNT_AUTOM_PAYMENTS   0.1095178647     5
```

Oracle Machine Learning

Machine Learning Notebooks included in Autonomous Databases

Key Features:

- Collaborative UI for data scientist and analysts
- Packaged with Autonomous Databases
- Quick start Example notebooks
- Easy access to shared notebooks, templates, permissions, scheduler, etc.
- OML4SQL
- OML4Py *coming soon*
- Supports deployment of OML models

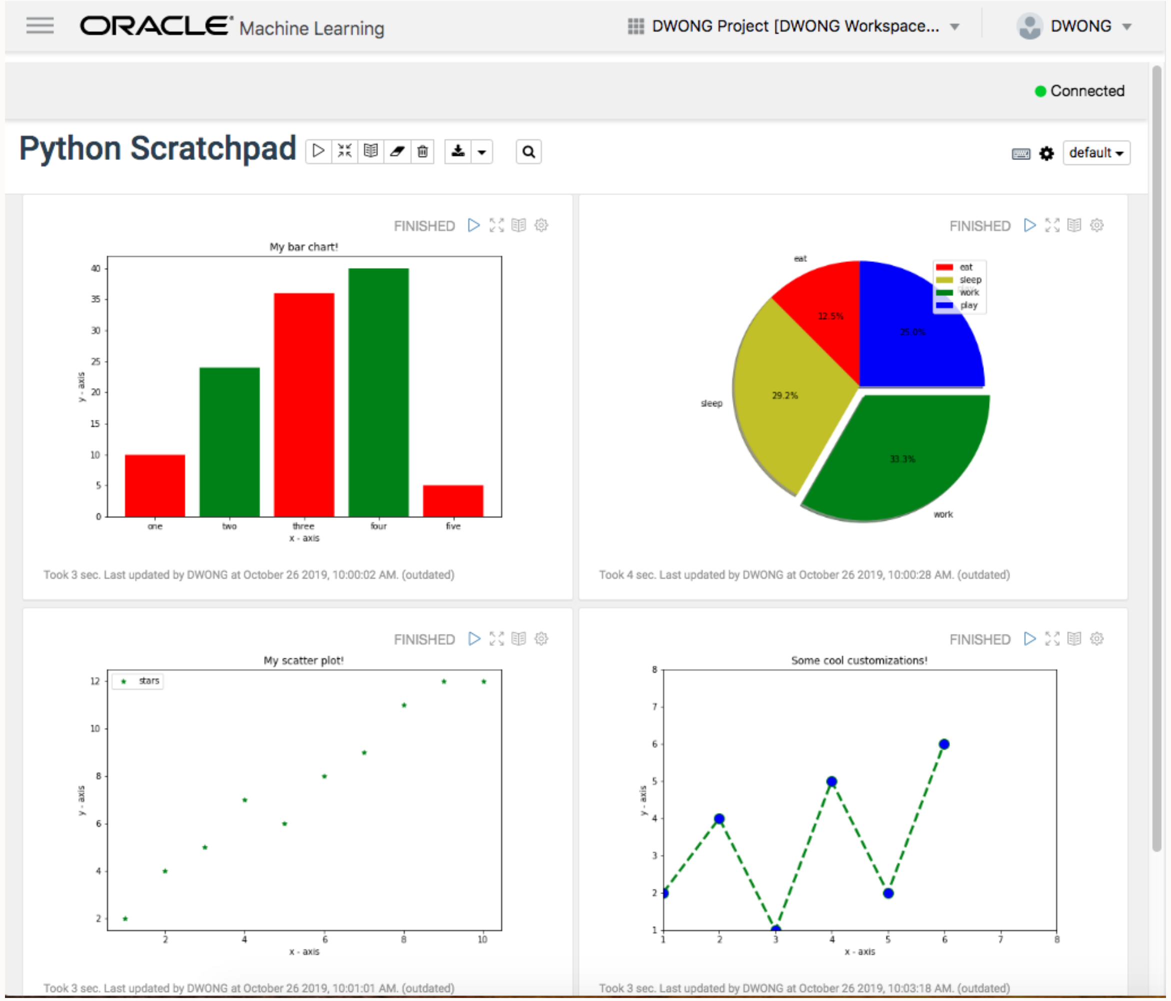


Oracle Machine Learning

Machine Learning Notebooks included in Autonomous Databases

Key Features:

- Collaborative UI for data scientist and analysts
- Packaged with Autonomous Databases
- Quick start Example notebooks
- Easy access to shared notebooks, templates, permissions, scheduler, etc.
- OML4SQL
- OML4Py *coming soon*
- Supports deployment of OML models



Oracle Machine Learning for R / Python

** Coming soon*

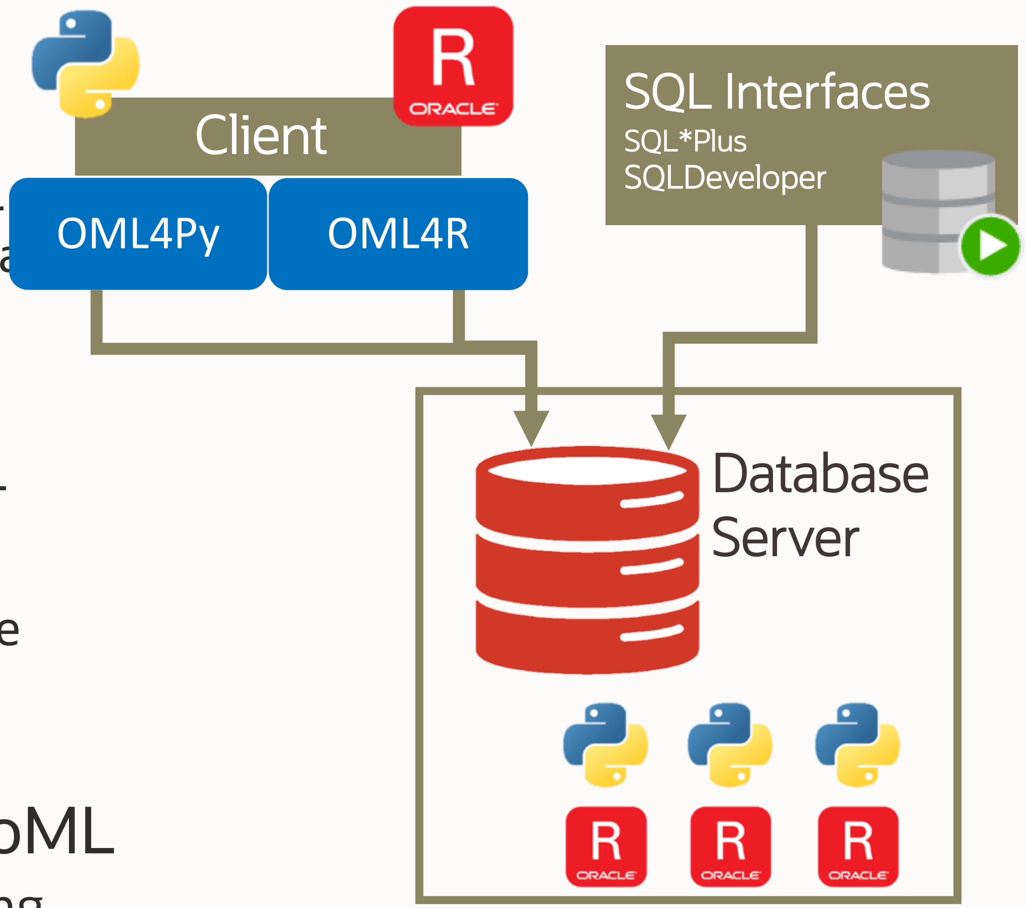
Multiple Components/APIs of Oracle Machine Learning

Transparency layer
- Leverage proxy objects so data remain in database
- Overload native functions translating functionality to SQL
- Use familiar R/Python syntax to manipulate database data

Parallel, distributed algorithms
- Scalability and performance
- Exposes in-database algorithms available from OML4SQL

Embedded execution
- Manage and invoke R or Python scripts in Oracle Database
- Data-parallel, task-parallel, and non-parallel execution
- Use open source packages to augment functionality

OML4Py, Automated Machine Learning - AutoML
- Feature selection, model selection, hyper-parameter tuning

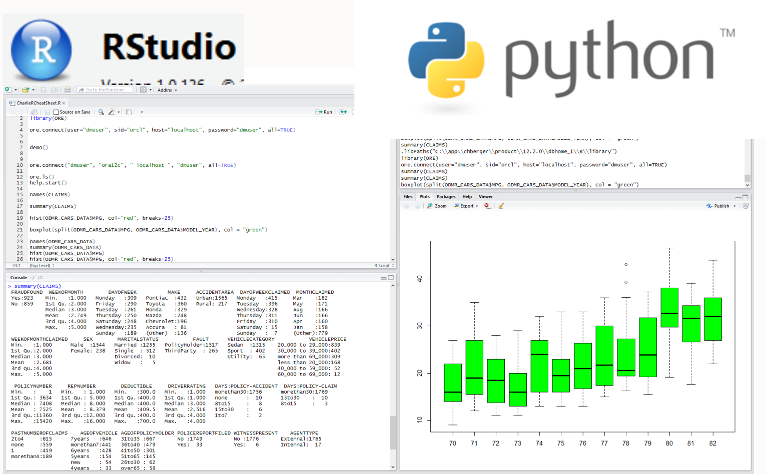


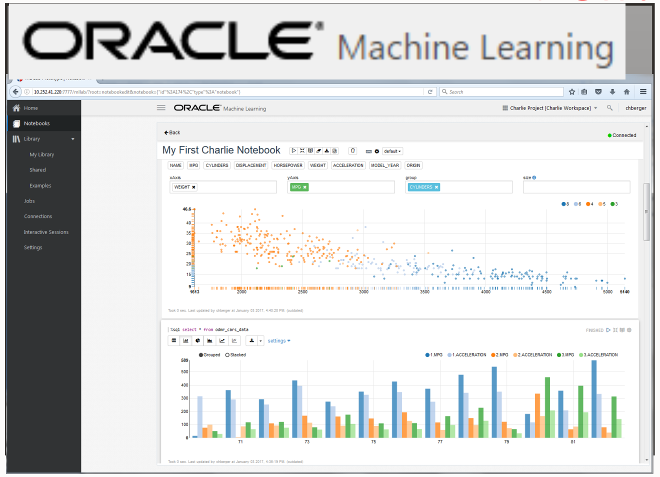
Oracle Machine Leaning

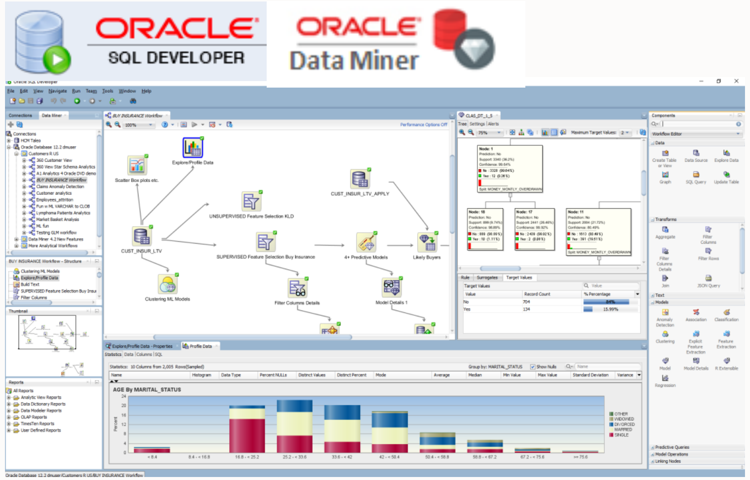
Multiple Languages UIs Supported for End Users & Apps Development

New!

New!



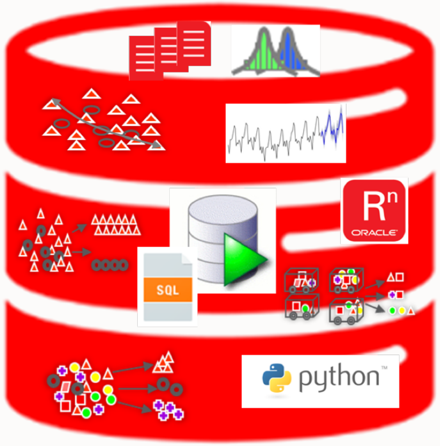




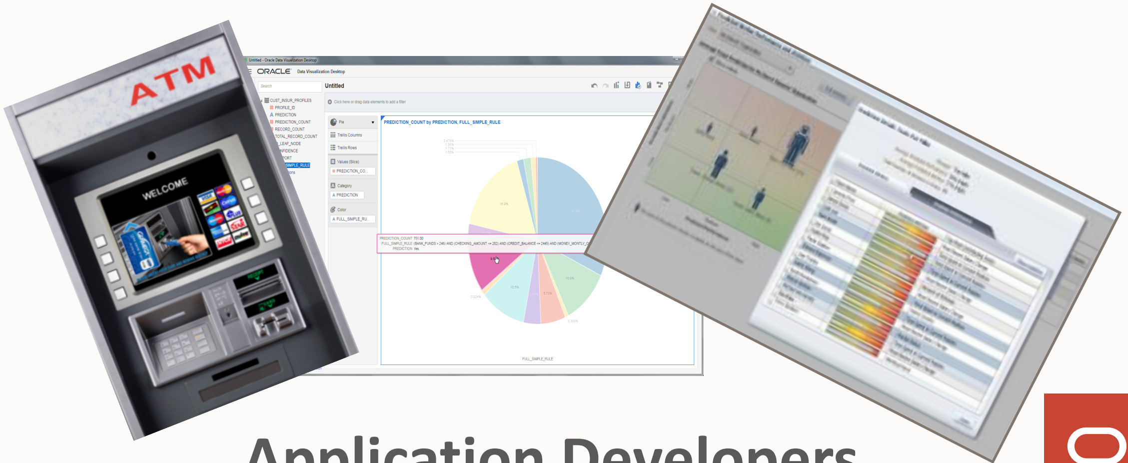
R & Python Data Scientists

Notebook Users & DS Teams

"Citizen" Data Scientists



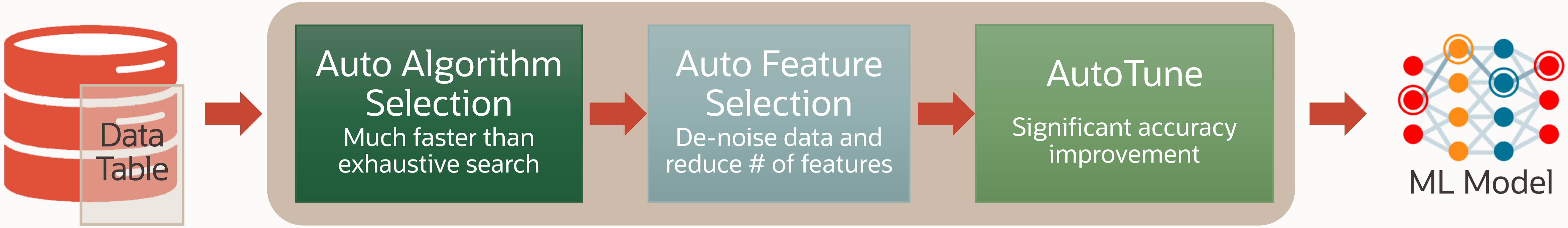
DBAs



Application Developers

Coming Soon! | AutoML – *new* with OML4Py

Increase data scientist productivity – reduce overall compute time



Auto Algorithm Selection

- Identify in-database algorithm that achieves highest model quality
- Find best algorithm faster than with exhaustive search

Auto Feature Selection

- **Reduce # of features by identifying most predictive**
- **Improve performance and accuracy**

Auto Tune Hyperparameters

- Significantly improve model accuracy
- Avoid manual or exhaustive search techniques

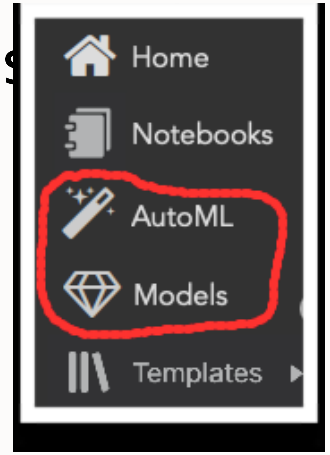
Enables non-expert users to leverage Machine Learning

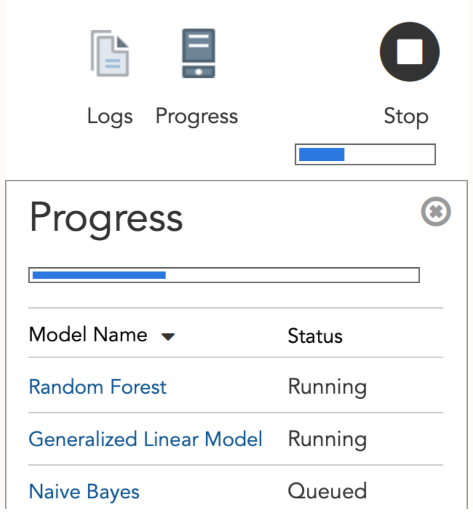
Coming Soon! | OML AutoML User Interface

"Code-free" user interface supporting automated end-to-end machine learning

Automate production and deployment of ML models

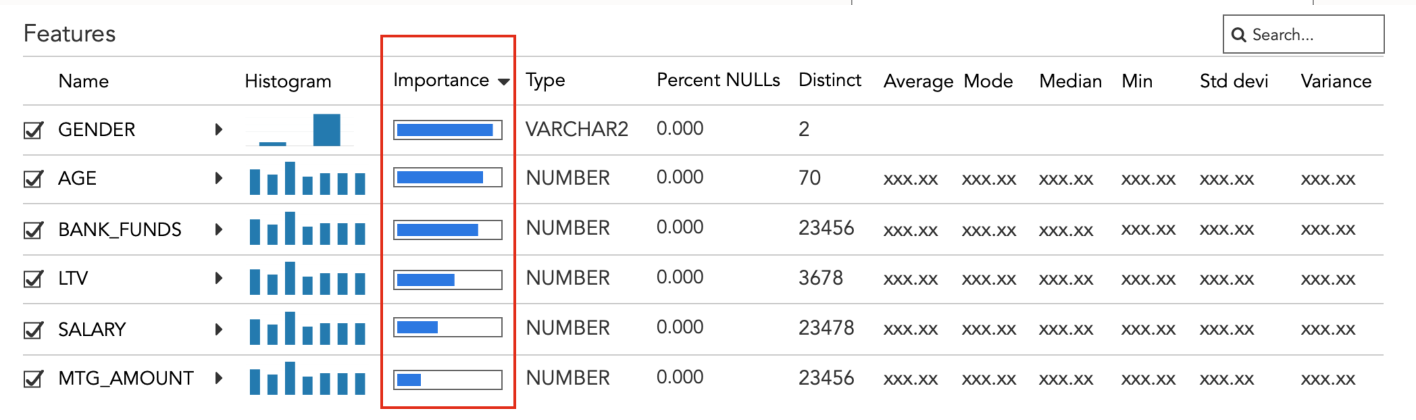
- Enhance Data Scientist productivity and user-experience
- Enable non-expert users to leverage ML
- Unify model deployment and monitoring
- Support model management





Features

- Minimal user input: data, target
- Model leaderboard
- Model deployment via REST
- Model monitoring
- Cognitive features for image and text



Features

Search...

Name	Histogram	Importance	Type	Percent NULLs	Distinct	Average	Mode	Median	Min	Std devi	Variance
GENDER			VARCHAR2	0.000	2						
AGE			NUMBER	0.000	70	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx
BANK_FUNDS			NUMBER	0.000	23456	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx
LTV			NUMBER	0.000	3678	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx
SALARY			NUMBER	0.000	23478	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx
MTG_AMOUNT			NUMBER	0.000	23456	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx	xxx.xx

Coming Soon! | OML AutoML User Interface

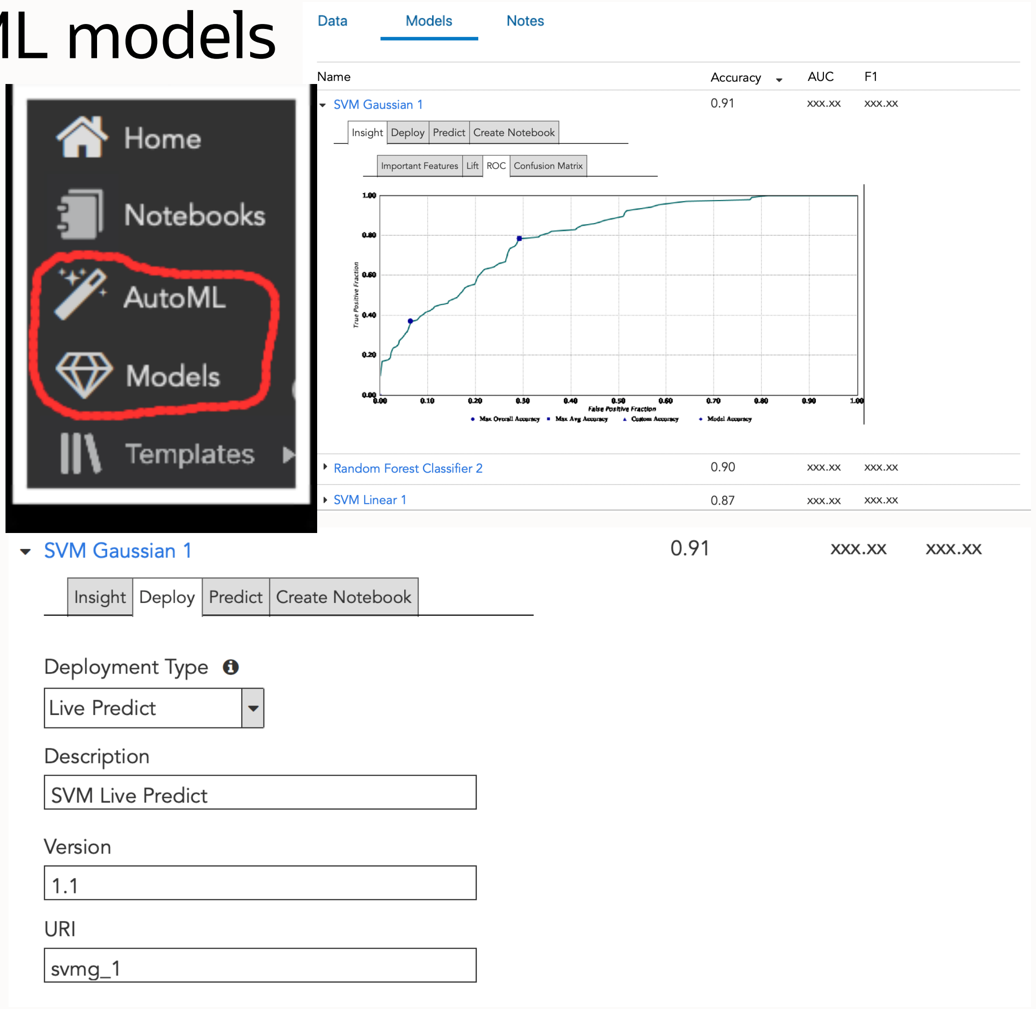
“Code-free” user interface supporting automated end-to-end machine learning

Automate production and deployment of ML models

- Enhance Data Scientist productivity and user-experience
- Enable non-expert users to leverage ML
- Unify model deployment and monitoring
- Support model management

Features

- Minimal user input: data, target
- Model leaderboard
- Model deployment via REST
- Model monitoring
- Cognitive features for image and text

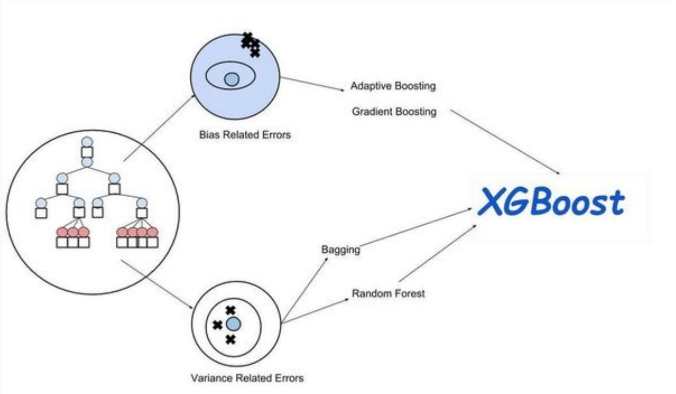


Coming Soon! | Algorithms for Database 20c

Two major new ML algorithms

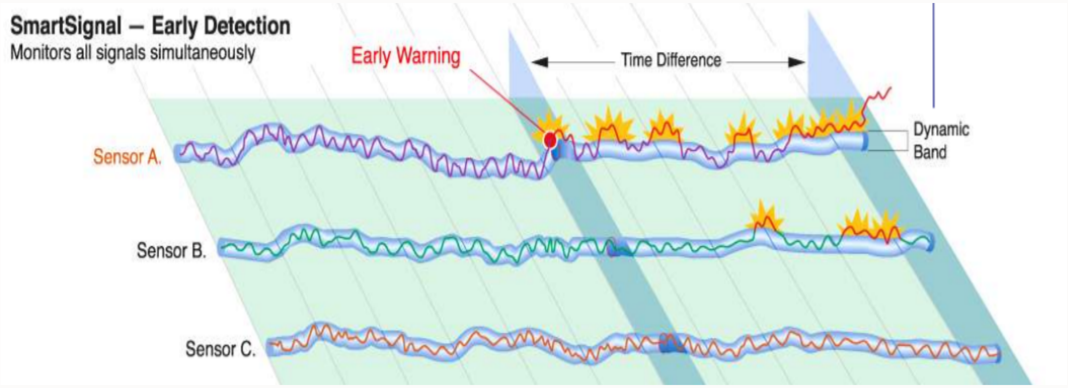
Gradient Boosted Trees (XGBoost)

- Highly popular and powerful algorithm – Kaggle winners
- Classification, regression, ranking, survival analysis



MSET-SPRT

- Multivariate State Estimation Technique - Sequential Probability Ratio Test (MSET-SPRT)
- Nonlinear, nonparametric anomaly detection algorithm designed to monitor critical processes.
- Detects subtle anomalies while also producing minimal false alarms.
- Calibrates expected behavior from historical normal operational sequence of monitored signals.
- Re-implemented and sped up in-DB and based on original Oracle Labs algorithm



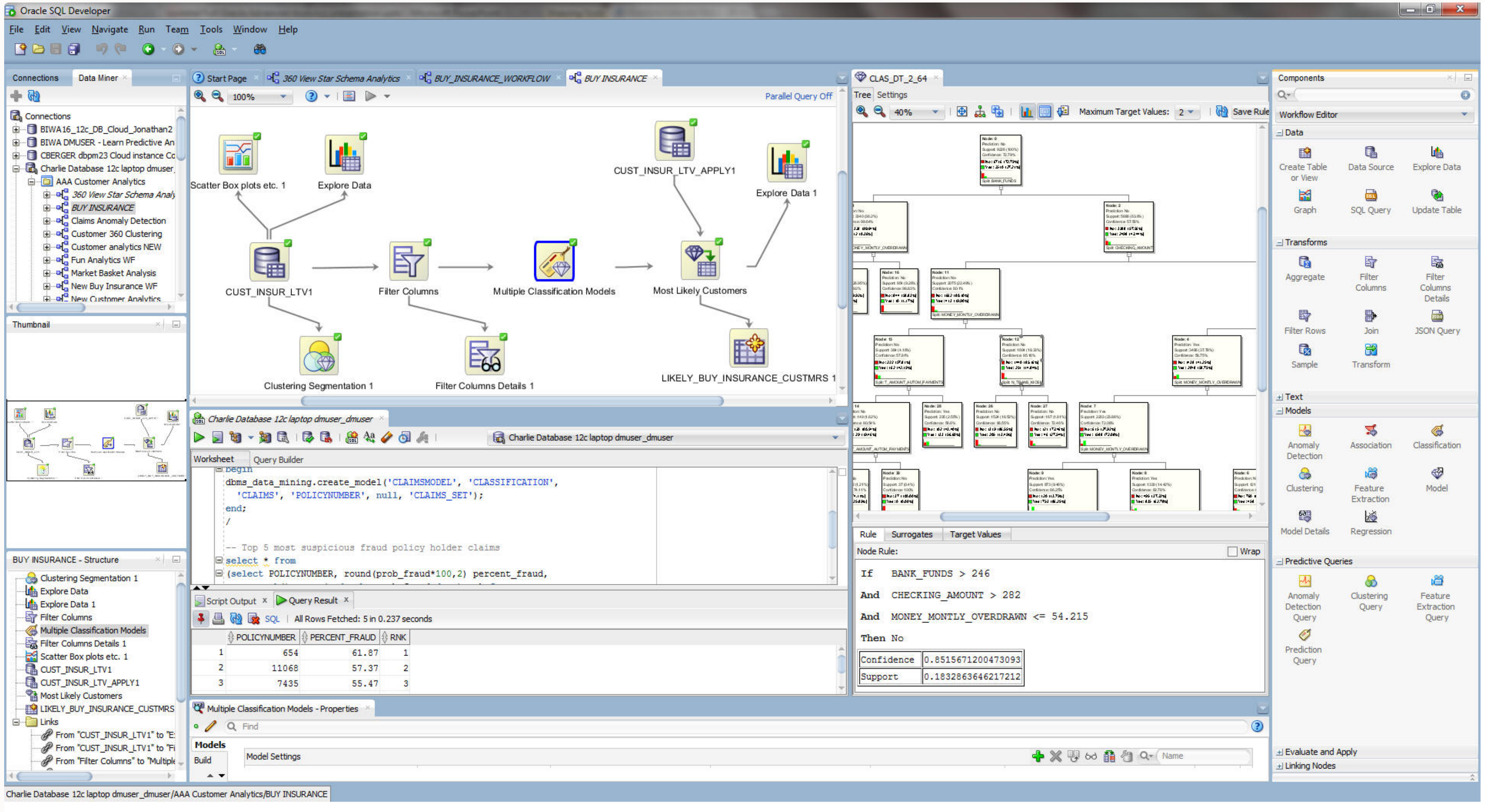
Oracle Data Miner UI

Drag and Drop, Workflows, Easy to Use UI for “Citizen Data Scientist”

Easy to use to define analytical methodologies that can be shared

SQL Developer Extension

Workflow API and generates SQL code for immediate deployment



Example Templates

New Notebook

Anomaly Detection

This notebook shows how to detect...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'Anomaly Detection' 'Machine...

1 Likes 24 0

Association Rules

Notebook to show the use of Assoc...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'SQL' 'Associations' 'Rules' 'M...

0 Likes 2 0

Attribute Importance

Notebook to identify key attributes...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'SQL' 'Attribute Importance' 'K...

0 Likes 10 0

Classification Prediction M...

Example notebook to predict custo...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'Classification' 'Prediction' 'De...

0 Likes 6 1

Clustering

This notebook shows how to identi...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'Clustering' 'K-Means' 'Expect...

0 Likes 9 0

My First Notebook

Oracle Machine Learning example ...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'SQL' 'Data' 'Graph'

0 Likes 15 0

Regression

This notebook shows how to predic...

Author:

Date Added: 5/4/18 6:59 AM

Tags: 'Regression' 'SVM' 'GLM' 'Logi...

0 Likes 6 0

Statistical Function

Oracle Machine Learning example ...

Author:

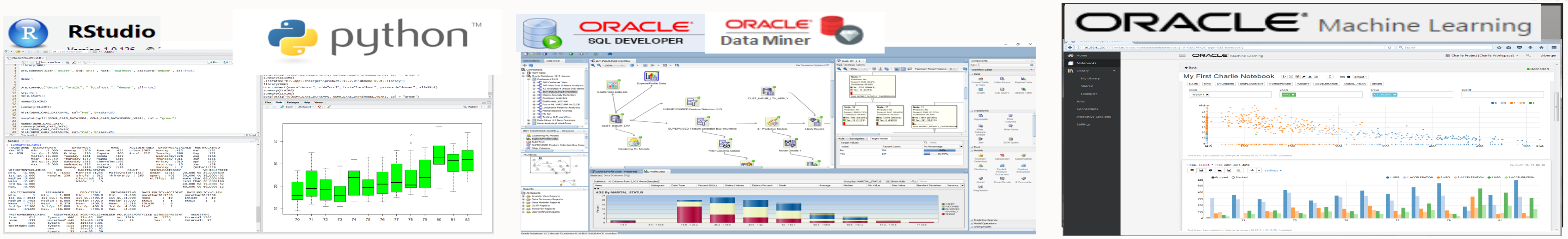
Date Added: 5/4/18 6:59 AM

Tags: 'Statistics' 'ANOVA' 'T-test' 'F-...

0 Likes 1 0

Manage and Analyze All Your Data

Architecturally, lots of options and flexibility



SQL / R / Python

Boil down the Data Lake

Big Data SQL / R

Object Store

NoSQL

ORACLE

kafka

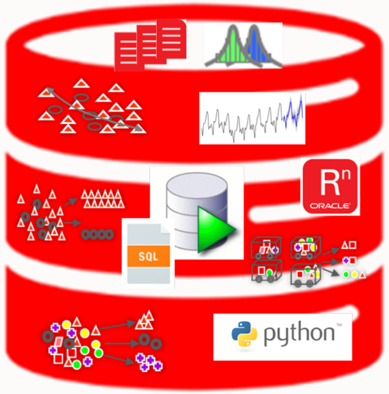
"Engineered Features"

– Derived attributes that reflect domain knowledge—key to best models e.g.:

- Counts
- Totals
- Changes over time

python

In-Database Machine Learning



More Models

Better Models

Faster, More Secure

Less Cost

No Need To Extract and Move Data	Data stays in Database	**Zero time required. No production impact.**
Data Preparation and Transformation	Accelerated with Automatic Data Prep	**No separate environment required. Much faster data prep. Data stays protected and secured.**
Data Mining and Model Building	SQL, R, Python Oracle Data Miner UI OML Notebooks	**Oracle Data Miner and AutoML greatly speed model building. Less skill required. No coding.**
No Need to Transform Production Data	Embedded Data Preparation	**No need for second production instance.**
Model Scoring	Accelerated Via Exadata Database Machine	**Faster model validation**
Ready to Deploy!	**Easy to repeat model building as often as needed**	

Oracle's Machine Learning & Adv. Analytics Algorithms

Classification
- Naïve Bayes
- Logistic Regression (GLM)
- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine
- Explicit Semantic Analysis

Clustering
- Hierarchical K-Means
- Hierarchical O-Cluster
- Expectation Maximization (EM)

Anomaly Detection
- One-Class SVM

Time Series
- State of the art forecasting using Exponential Smoothing
- Includes all popular models e.g. Holt-Winters with trends, seasons, irregularity, missing data

Regression
- Linear Model
- Generalized Linear Model
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- LASSO *

Attribute Importance
- Minimum Description Length
- Principal Comp Analysis (PCA)
- Unsupervised Pair-wise KL Div
- CUR decomposition for row & AI

Association Rules
- A priori/ market basket

Predictive Queries
- Predict, cluster, detect, features

SQL Analytics
- SQL Windows, SQL Patterns, SQL Aggregates

Feature Extraction
- Principal Comp Analysis (PCA)
- Non-negative Matrix Factorization
- Singular Value Decomposition (SVD)
- Explicit Semantic Analysis (ESA)

Text Mining Support
- Algorithms support text
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA) for document similarity

Statistical Functions
- Basic statistics: min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.

R Packages
- CRAN R Algorithm Packages through Embedded R Execution
- Spark MLlib algorithm integration

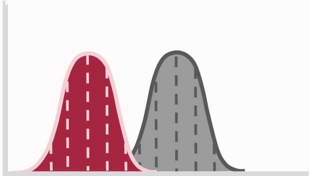
Exportable ML Models
- REST APIs for deployment

• OAA (Oracle Data Mining + Oracle R Enterprise) and ORAAH combined
• OAA includes support for Partitioned Models, Transactional, Unstructured, Geo-spatial, Graph data. etc,

Oracle's Machine Learning & Adv. Analytics Algorithms

Statistical Functions

- Descriptive statistics (e.g. median, stdev, mode, sum, etc.)
- Hypothesis testing (t-test, F-test, Kolmogorov-Smirnov test, Mann Whitney test, Wilcoxon Signed Ranks test
- Correlations analysis (parametric and nonparametric e.g. Pearson's test for correlation, Spearman's rho coefficient, Kendall's tau-b correlation coefficient)
- Ranking functions
- Cross Tabulations with Chi-square statistics|



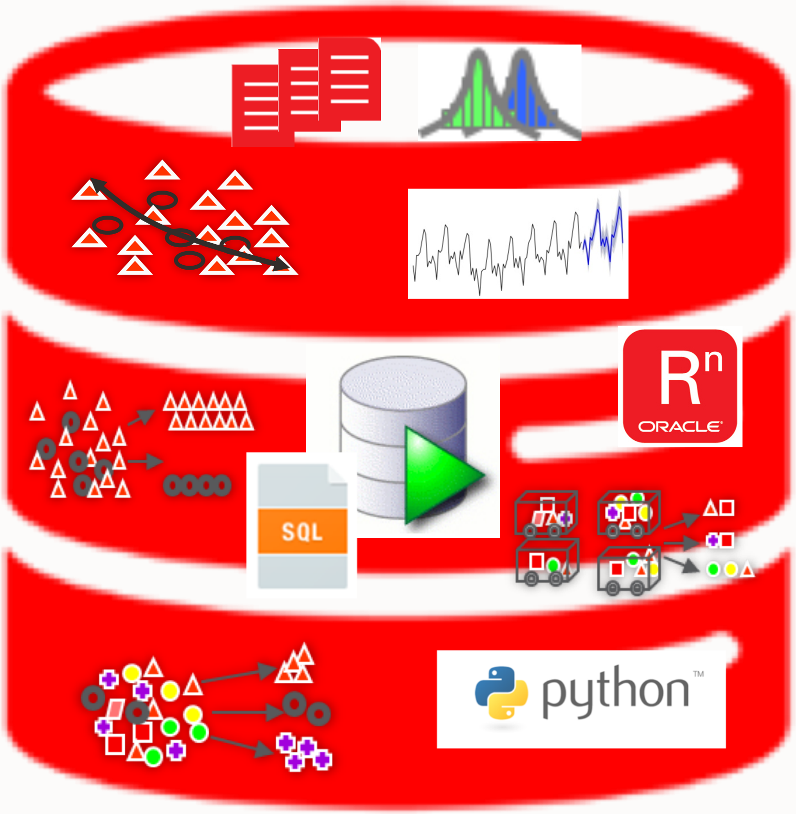
- Linear regression
- ANOVA (Analysis of variance)
- Test Distribution fit (e.g. Normal distribution test, Binomial test, Weibull test, Uniform test, Exponential test, Poisson test, etc.)
- Statistical Aggregates (min, max, mean, median, stdev, mode, quantiles, plus x sigma, minus x sigma, top n outliers, bottom n outliers)

Analytical SQL

- SQL Windows
- SQL Aggregate functions
- LAG/LEAD functions
- SQL for Pattern Matching
- Additional approximate query processing: APPROX_COUNT, APPROX_SUM, APPROX_RANK
- Regular Expressions

ML and AI are just "Algorithms"

Algorithms Operate on Data



Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Move the Algorithms; Not the Data!;
It Changes *Everything*!

Thank You

Any Questions ?

Sandesh Rao
VP AIOps for the Autonomous Database

@sandeshr
https://www.linkedin.com/in/raosandesh/
https://www.slideshare.net/SandeshRao4

ORACLE