INTRODUCTION TO MACHINE LEARNING AND ALGORITHMS

ABSTRACT

Machine learning addresses the vital question of how to construct computers that improve by itself through experience and self-learning. It is one of today's most rapidly and fast growing technical fields, lying at the intersection of computing and statistics, and at the core of computing and information science. Recent progress and development in machine learning has been driven each by the event of latest learning algorithms and theory and by the continuing explosion within the handiness of on-line information and affordable computation. The self- learning of data-intensive machine-learning ways are often found throughout science, technology and commerce, resulting in a lot of evidence-based decision-making across several walks of life, together with health care, producing, education, money modeling, policing, promoting and plenty of alternative fields.

INTRODUCTION

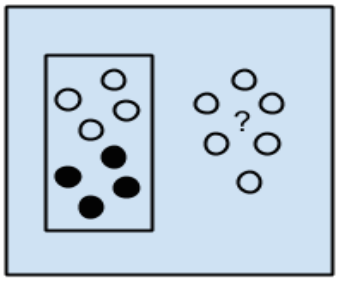
Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information. While many machine learning algorithms have been there for a long time, the ability to automatically implement complex mathematical calculations to big data – over and over again, faster is a recent development. When thinking of Machine learning on a high level, which is often represented as just a mystery. Even if we do not go to the details of how it works, we should have a clear idea of what result we want to get.

So, when we think of the problem we want to solve, it makes sense to apply machine learning when there is a repetitive work which has a clear outcome but does not conform to standard clearly defined rules.

1. Algorithms Grouped by Learning Style

There are different methods by which algorithm can solve a problem based on its interaction with the experience it has gained or environment call the input data. It is popular in machine learning and artificial intelligence books to first consider the learning styles that an algorithm can adopt. There are only a few main machines learning styles or learning models that an algorithm can adopt and we will go through them here with a examples of algorithms and type of problems that suits them. This way of organizing algorithms is useful because it forces us to think about the roles of the input data and the model preparation process and select one that is the most appropriate for problem in order to get the best results out of it.

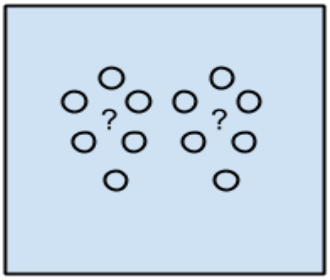
1.1 Supervised Learning



Supervised Learning Algorithms

Input data is called learning data and has a known result such as spam or not-spam or a stock price at a time. A model is prepared through a training process in which it is required to make predictions and it is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Example problems are classification and regression. Example algorithms include Logistic Regression and the Back Propagation Neural Network.

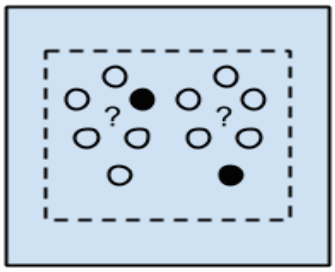
1.2 Unsupervised Learning



Unsupervised Learning Algorithms

Input data is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity. Example problems are clustering, dimensionality reduction and association rule learning. Example algorithms include: the Apriori algorithm and k-Means.

1.3 Semi-Supervised Learning



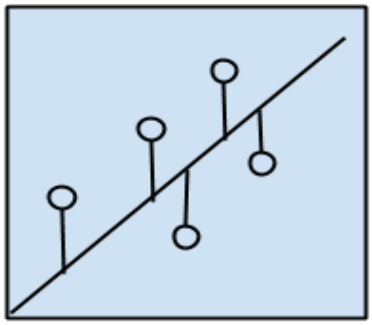
Semi-supervised
Learning Algorithms

Input data is a mixture of labeled and un labelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions. Example problems are classification and regression. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabeled data.

2. Algorithms Grouped by Similarity

Algorithms are often grouped by similarity in terms of their function. For example, tree-based methods, and neural network inspired methods. This is the most useful way to group algorithms and it is the approach we will use here. This is a useful grouping method, but is not perfect. There are still algorithms that could just as easily fit into multiple categories like Learning Vector Quantization that is both a neural network inspired method and an instance-based method. There are also categories that have the same name that describe the problem and the class of algorithm such as Regression and Clustering.

2.1 Regression Algorithms



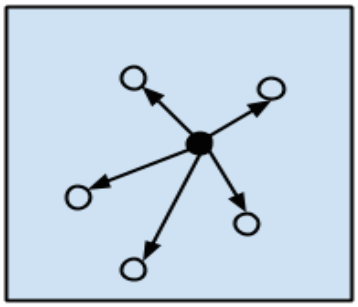
Regression Algorithms

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning. This may be confusing because we can use regression to refer to the class of problem and the class of algorithm. Really, regression is a process. The most popular regression algorithms are as follows:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)

- Locally Estimated Scatterplot Smoothing (LOESS)

2.2 Instance-based Algorithms

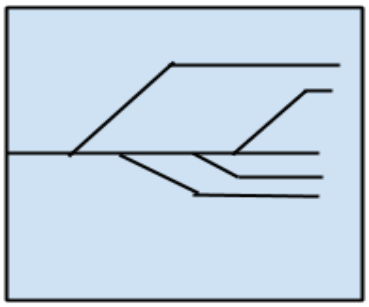


Instance-based Algorithms

Instance-based learning model is a decision problem with instances or examples of training data that are deemed important or required to the model. Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances. The most popular instance-based algorithms are:

- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

2.3 Regularization Algorithms

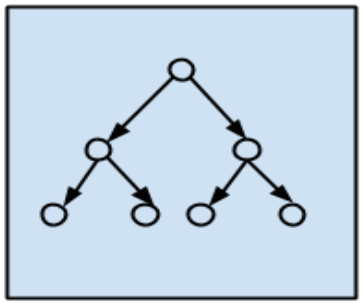


Regularization Algorithms

An extension made to another method that penalizes models based on their complexity, favoring simpler models that are also better at generalizing. It is listed regularization algorithms separately here because they are popular, powerful and generally simple modifications made to other methods. The most popular regularization algorithms are:

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)

2.4 Decision Tree Algorithms

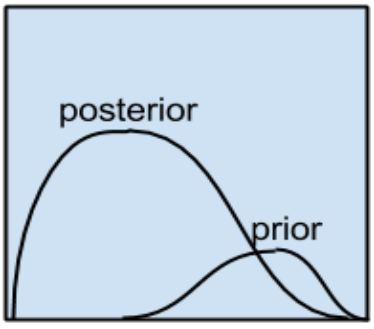


Decision Tree Algorithms

Decision tree methods construct a model of decisions made based on actual values of attributes in the data. Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning. The most popular decision tree algorithms are:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

2.5 Bayesian Algorithms



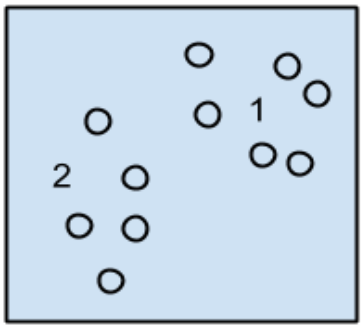
Bayesian Algorithms

Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification and regression. The most popular Bayesian algorithms are:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)

- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

2.6 Clustering Algorithms

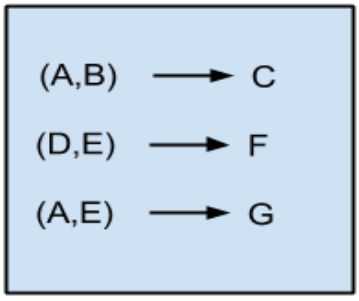


Clustering Algorithms

Clustering, like regression, describes the class of problem and the class of methods. Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality. The most popular clustering algorithms are:

- k-Means
- k-Medians
- Expectation Maximization (EM)
- Hierarchical Clustering

2.7 Association Rule Learning Algorithms

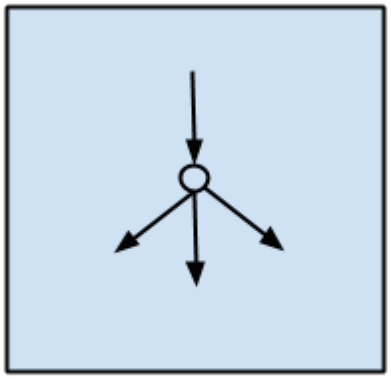


Association Rule
Learning Algorithms

Association rule learning methods extract rules that best explain observed relationships between variables in data. These rules can discover important and commercially useful associations in large multidimensional datasets that can be exploited by an organization. The most popular association rule learning algorithms are:

- Apriori algorithm
- Eclat algorithm

2.8 Artificial Neural Network Algorithms

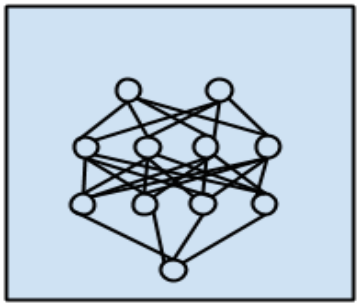


Artificial Neural Network Algorithms

Artificial Neural Networks are models that are inspired by the structure or function of biological neural networks. They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types. The most popular artificial neural network algorithms are:

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)

2.9 Deep Learning Algorithms



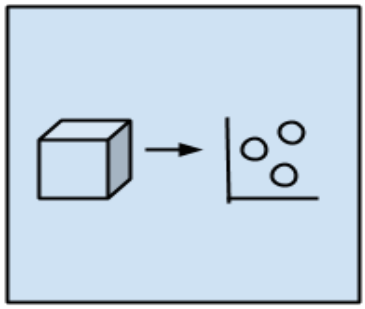
Deep Learning Algorithms

Deep Learning methods are a modern update to Artificial Neural Networks that exploit abundant cheap computation. They are concerned with building much larger and more complex neural networks and, as commented on above, many methods are concerned with semi-supervised learning problems where large datasets contain very little labeled data. The most popular deep learning algorithms are:

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)

- Stacked Auto-Encoders

2.10 Dimensionality Reduction Algorithms

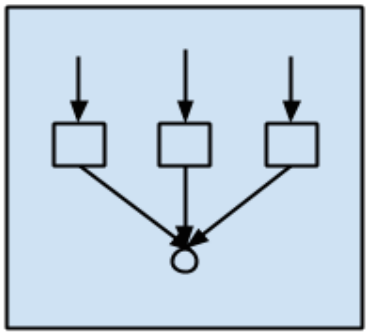


Dimensional Reduction Algorithms

Like clustering methods, dimensionality reduction seeks and exploit the inherent structure in the data, but in this case in an unsupervised manner or order to summarize or describe data using less information. This can be useful to visualize dimensional data or to simplify data which can then be used in a supervised learning method. Many of these methods can be adapted for use in classification and regression.

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)

2.11 Ensemble Algorithms



Ensemble Algorithms

Ensemble methods are models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction. Much effort is put into what types of weak learners to combine and the ways in which to combine them. This is a very powerful class of techniques and as such is very popular.

- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (blending)
- Gradient Boosting Machines (GBM)
- Gradient Boosted Regression Trees (GBRT)
- Random Forest

Machine Learning research has been extremely active the last few years. The result is a large number of very accurate and efficient algorithms that are quite easy to use for a practitioner. It seems rewarding and almost mandatory for (computer) scientist and engineers to learn how and where Machine Learning can help to automate tasks 4 or provide predictions where humans have difficulties to comprehend large amounts of data

SAP HANA (MACHINE LEARNIING)

SAP HANA (High Performance Analytic Appliance) provides native in-memory capabilities for predictive analytics and machine learning corporal punishment at new speeds directly within the information. together with different advanced process capabilities like spacial, text or graph analysis SAP HANA inherently permits multi-modal analysis eventualities.

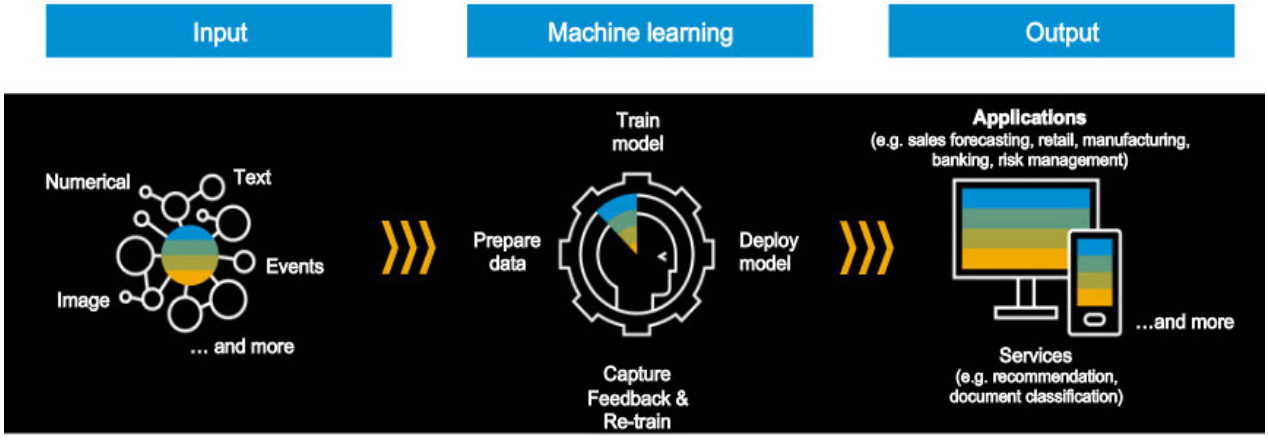
The SAP HANA predictive analytics library (PAL) includes over ninety algorithms for eventualities like cluster analysis, outlier detection, classification and multivariate analysis, association analysis, link prediction and recommendation analysis at the side of several applied mathematics and information preparation algorithms.

The use of the PAL permits you to bring and apply predictive and machine learning algorithms to wherever the info is keep in distinction to standalone, dedicated machine learning platforms, wherever information 1st should be derived and removed of the information so as to realize and apply any price from advanced analytics. you'll then execute predictive analysis or apply predictions with trained models as a district of any information transactions, co-located along with your application database-layer like S/4HANA, biological attack on HANA or the other SAP or custom application running on SAP HANA. With the utilization of PAL, you'll build and run smarter applications directly on SAP HANA.

The SAP HANA platform's distinctive ability to use native, aboard execution engines suggests that you'll perform predictive calculations for coaching, validating, and grading while not information extraction or maybe manipulation. Running predictive models in-memory ends up in dramatic production enhancements additionally to a more-efficient use of system resources. If you'll build quicker selections supported period analysis of information, you'll usually build higher and more-informed selections. once these execution engines square measure combined with the event-processing capabilities of the platform, you'll be instantly alerted and take action as before long as a state of affairs or chance is detected. think about a state of affairs wherever your victimization numerous client records in Associate in Nursing analysis to work out clusters of enticing market segments among the population. process such an oversized range of records is already an upscale and long task. However, extracting and transferring these records to a separate analytics server conjointly creates further employment and price for the remainder of the IT infrastructure. the power of SAP HANA to perform predictive calculations directly from among SQL script permits you to use existing SQL experience in your organization whereas providing on-the-fly process on compressed information with unmatched speed and potency.

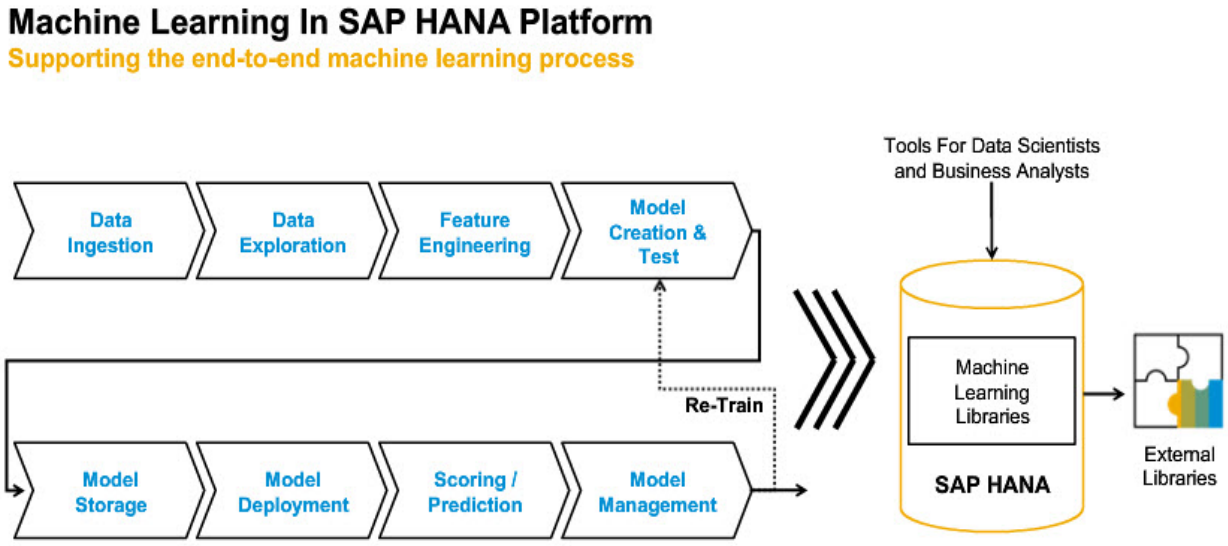
.

Machine Learning – What it is and how does it work?



In Machine learning algorithm encompasses decision making and prediction process decouple the logic from algorithm. Increases the flexibility of applications. Input data, train model, test that model is authentic. new knowledge comes retrain the model to require advantage. Model becomes additional sturdy and arduous. 3 phases this is often

input, machine learning, output. Input may well be text, images, knowledge cleansing, build model, set knowledge to coach model. The model manager tool at intervals SAP predictive analytics provides automatic performance-tuning capabilities to assist make sure that models square measure tuned to be operational at peak performance in the slightest degree times for optimum outcomes. The tool options a browser-based, single-sign-on setting and easy programing interface designed for knowledge analysts. once it's combined with alternative elements from SAP predictive Analytics, SAP HANA becomes one platform for all predictive workflows and automates the total prophetic lifecycle from model creation to readying and even in progress model validation. The IT landscape becomes less complicated, and users have one analytics platform in spite of the predictive technology they use.

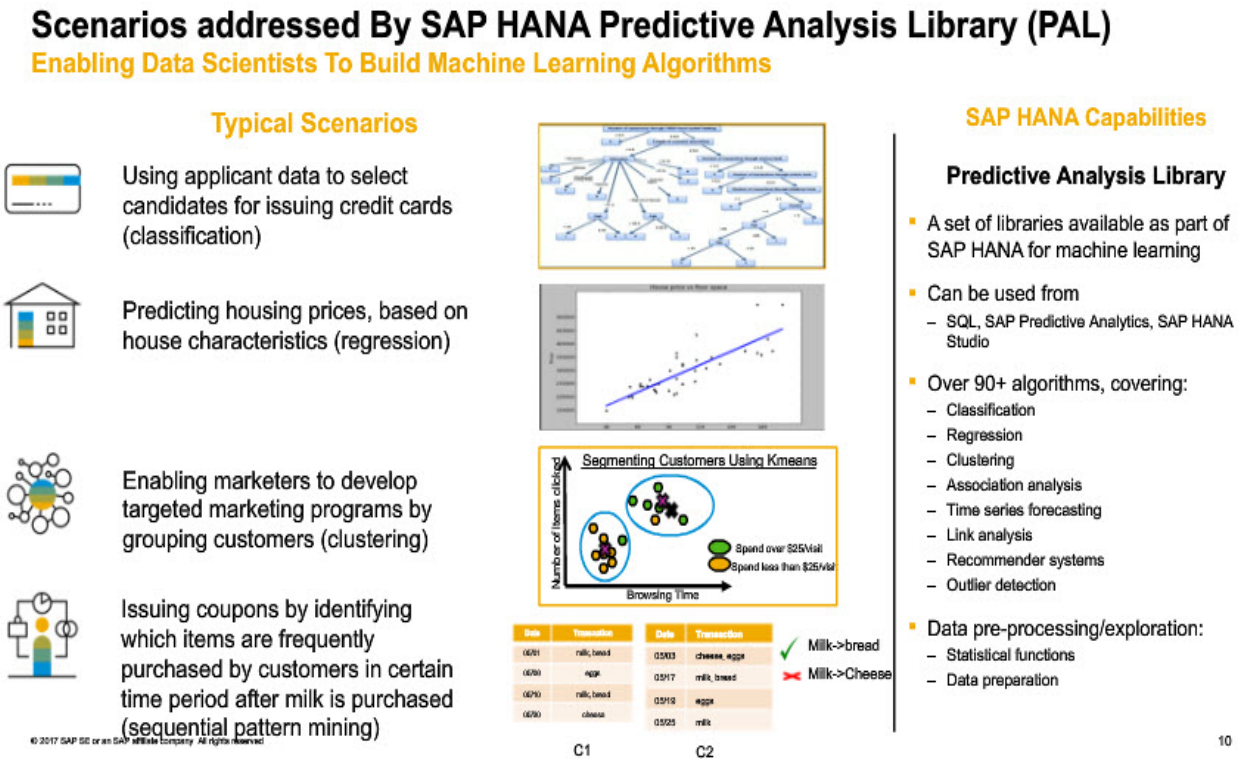


Embedded model in application of SAP HANA.

Machine learning in SAP HANA is an end to end process. It starts with data ingestion. HANA can load data streaming. After this it is data exploration. It tells us how does data looks? Does it have any missing data? It also provides us with tools to help. It has the feature of engineering to transform input itself. Once data is ready, split in two groups. One is for coaching, alternative one for testing and validate cubic centimeter formula. It stores model in HANA info. Models ought to be deployed during a form of alternative ways i.e. SQL or as a service. The grading or prediction happens in real time, execute models quickly. Model management ensures that you just will retrain models on an occasion driven basis or periodic basis.

SAP HANA Predictive Analysis Library (PAL)

The predictive analysis library, or PAL, is meant to require advantage of the flexibility of SAP HANA to host execution engines and perform native calculations in memory. in contrast to the previous choice that uses AN external predictive server for process, this SAP HANA–native library permits users to perform in-database data processing and applied math calculations with glorious performance on giant information sets. While the PAL cannot replicate all 5,800 algorithms that are offered with R, the PAL contains the SAP native C++ implementations of the foremost ordinarily used algorithms. the quantity of algorithms supported in PAL has been growing with each service pack of SAP HANA.



The figures show the situations self-addressed by PAL. The figures show candidates appropriate for master card process and its classification situation, credit score, financial gain level, earth science exploitation call tree. It additionally tells United States that one can default.

Builds model on historic knowledge, new candidates to predict if they're default or to increase credit limit.

Regression model for predicting house costs, build model, sporadically trigger retrain model as required to confirm model sure thing is correct.

Look at customers to run selling to cluster as a logical entity bunch, k-means, you'll need to place some selling programs. To analyze client group action knowledge i.e. customers UN agency bought milk, did they obtain, exploitation successive pattern mining rule.

There are more than 90 algorithms in PAL. The list of algorithms is given in the figure below. Data scientists are still free to use an external R server for algorithms or functions not included in the PAL. This enables any combination of algorithms to help ensure that users are benefiting from the speed of SAP HANA as much as possible without giving up the flexibility and extensibility of algorithms in R.

New and enhanced algorithms in SAP HANA 2 SPS01

Machine Learning Algorithms – 90+ and growing

* New in HANA 2 SPS0
** New in HANA 2 SPS01

Classification Analysis
- CART
- C4.5 Decision Tree Analysis
- CHAID Decision Tree Analysis
- K Nearest Neighbour
- Logistic Regression Elastic Net
- Back-Propagation (Neural Network)
- Naïve Bayes
- Support Vector Machine
- Random Forests
- Gradient Boosting Decision Tree (GBDT)*
- Linear Discriminant Analysis (LDA)*
- Confusion Matrix
- Area Under Curve (AUC)
- Parameter Selection / Model Evaluation

Regression
- Multiple Linear Regression Elastic Net
- Polynomial, Exponential, Bi-Variate Geometric, Bi-Variate Logarithmic Regression
- Generalized Linear Model (GLM)*
- Cox Proportional Hazards Model*

Cluster Analysis
- ABC Classification
- DBSCAN
- K-Means/ Accelerated K-Means**
- K-Medoid Clustering
- K-Medians
- Kohonen Self Organized Maps
- Agglomerate Hierarchical
- Affinity Propagation
- Latent Dirichlet Allocation (LDA)
- Gaussian Mixture Model (GMM)
- Cluster Assignment

Time Series Analysis
- Single/Double/ Brown /Triple Exp.Smoothing
- Forecast Smoothing
- Auto - ARIMA/ Seasonal ARIMA
- Croston Method
- Forecast Accuracy Measure
- Linear Regression with Damped Trend and Seasonal Adjust
- Test for White Noise, Trend, Seasonality
- Fast Fourier Transform (FFT)*
- Correlation Function*

Association Analysis
- Apriori, Apriori Lite
- FP-Growth
- KORD – Top K Rule Discovery
- Sequential Pattern Mining*

Probability Distribution
- Distribution Fit/ Weibull analysis
- Cumulative Distribution Function
- Quantile Function
- Kaplan-Meier Survival Analysis

Outlier Detection
- Inter-Quartile Range Test (Tukey's Test)
- Variance Test
- Anomaly Detection
- Grubbs Outlier Test

Recommender Systems
- Factorized Polynomial Regression Models**

Link Prediction
- Common Neighbors, Jaccard's Coefficient, Adamic/Adar, Katzβ

Statistical Functions
- Mean, Median, Variance, Standard Deviation, Kurtosis, Skewness
- Covariance Matrix
- Pearson Correlations Matrix
- Chi-squared Tests:
  - Test of Quality of Fit
  - Test of Independence
- F-test (variance equal test)
- Data Summary*
- ANOVA**
- One-sample Median Test**
- T Test**
- Wilcox Signed Rank Test**

Data Preparation
- Sampling, Binning, Scaling, Partitioning
- Principal Component Analysis (PCA) / PCA Projection

Other
- Weighted Scores Table
- Substitute Missing Values

SAP Predictive Analytics provides a graphical expert mode to use PAL algorithms and supports the intermixing of PAL and R algorithms within the same workflow.

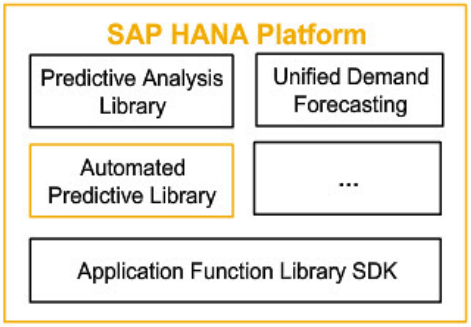
SAP Automated Predictive Library (APL)

SAP Automated Predictive Library (APL)

Enabling business analysts and data scientists to build predictive models quickly

Typical Scenarios

- Similar use cases as described before can be implemented quickly. For example: credit card application processing, marketing segmentation, house price estimation, etc.
- Automated analytics libraries can be used by business analysts and data scientists, without requiring extensive machine learning expertise

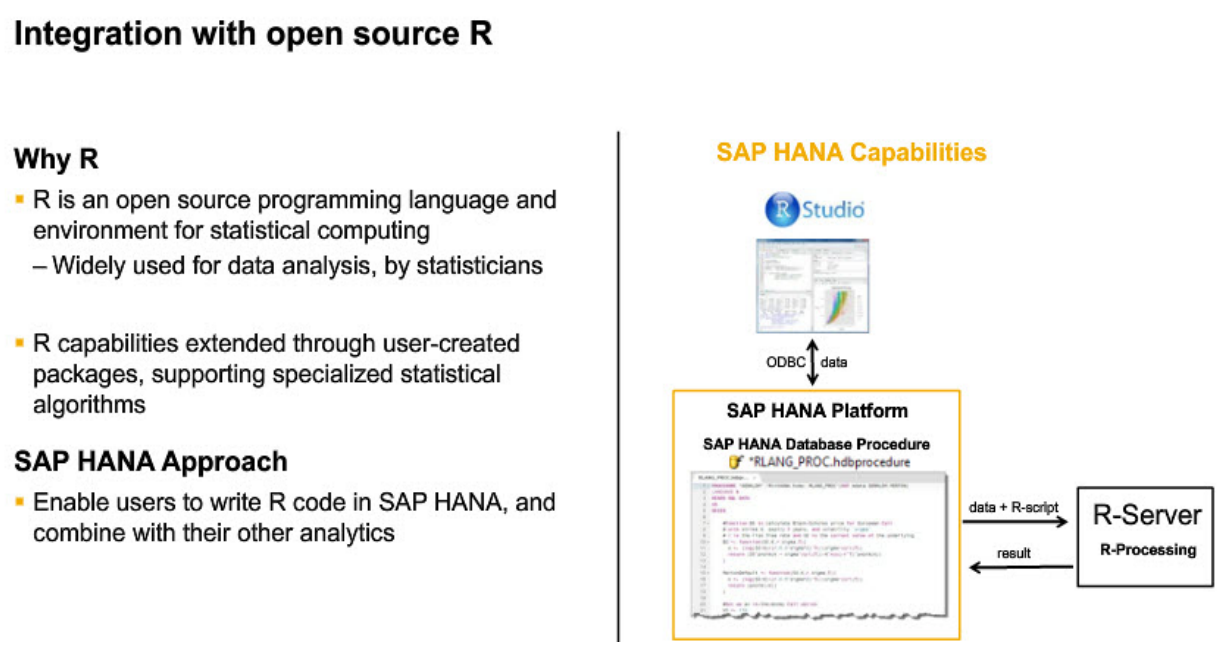


- Embedded inside SAP HANA
- Models provided
  - Classification/regression models
  - Clustering models
  - Time series analysis models
  - Association rules
  - Recommendation models
  - Social network analysis models

The automated analytics interface within SAP Predictive Analytics provides data analysts and data scientists with automated machine learning capabilities and can create predictive models without requiring data science experience. It also does not require a complex predictive model as input, it simply needs to be configured and told what type of determining function needs to be applied to the data. If you are not in a data scientist role, you are typically analyzing data to solve specific problems related to your job. An automated machine learning system enables you to focus on the business problem you are trying to solve instead of algorithmic selection, model creation, and other predictive workflows. With SAP Predictive Analytics, this process is completely automated and therefore puts the following capabilities in the hands of almost all business users, whether they consider themselves analysts or scientists:

• Classification and regression models
• Clustering models
• Time-series analysis models
• Recommendation models

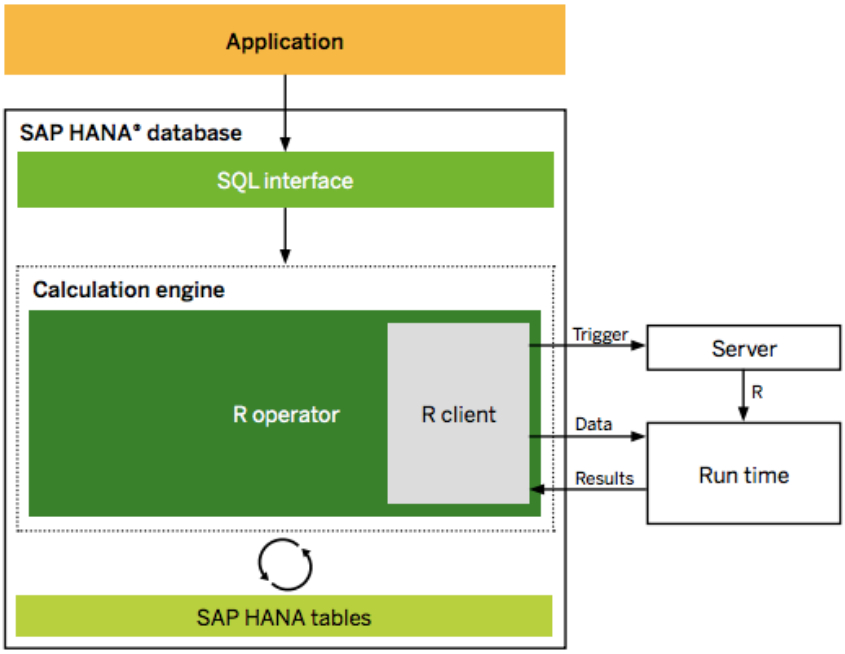
Predictive Capabilities of SAP HANA: R



Data scientists usually produce predictive models on historical information sets by applying one or a lot of mathematical algorithms to capture the implicit relationships at intervals the information. Use of those algorithms needs a big understanding of the information and a solid grasp of statistics and alternative mathematical ideas. it's not uncommon for a knowledge somebody to pay days or weeks analyzing the information before making a strong and stable predictive model. In some cases, the information somebody might even have to be compelled to produce his or her own algorithms to resolve more-complex issues or produce a predictive model that's specific to associate degree trade. SAP HANA supports multiple ways for victimization information science–specific algorithms within the predictive method. the subsequent section discusses the advantages and challenges of every.

The ASCII text file language R is that the most well-liked predictive-modeling setting within the world. As a language created by mathematicians, it had been designed from the start to be simply extensile and supports individuals sharing what they need written with others to use in their own modeling tasks. At the time of this writing, there square measure quite five,800 ASCII text file algorithms publically offered and innumerable others that square measure thought-about the proprietary material possession of their creators. SAP predictive analytics offers associate

degree skilled mode to desktop users to allow you to use any combination of algorithms from a range of libraries. Libraries embody those in SAP HANA like the PAL and APL, further as associate degree external R server.

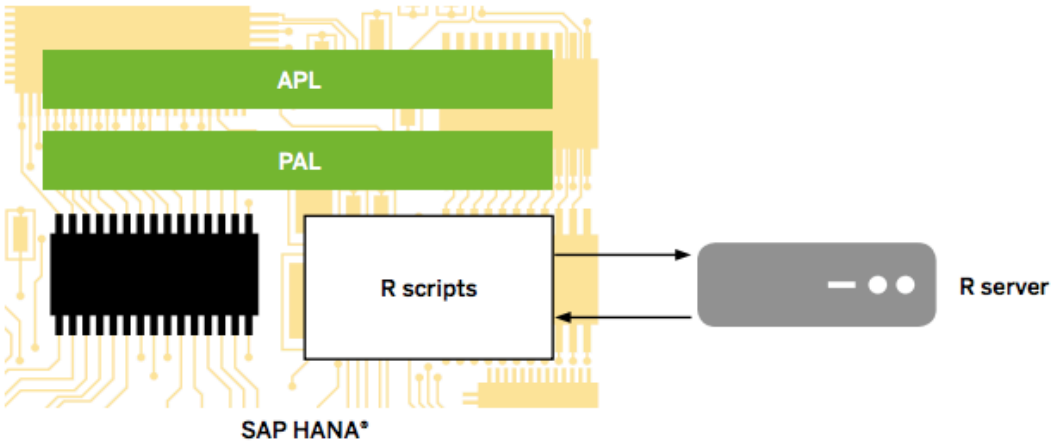


SAP HANA supports this sort of sidecar preparation to supply information scientists the final word flexibility to use any rule they require or perhaps produce their own in R. They decision R scripts through SQL script in SAP HANA then pass the scripts to associate degree external R server, in conjunction with all of the desired computer file. The results square measure sent back to SAP HANA and combined with any native information needed to complete the question within the figure higher than. However, desegregation the results from discretional scripting code employing a lead to important overhead needed to extract and transfer information to the R server for process. activity all calculations with an area in-memory predictive engine, like the PAL for SAP HANA, avoids this overhead.

Comparing R, PAL and APL Capabilities

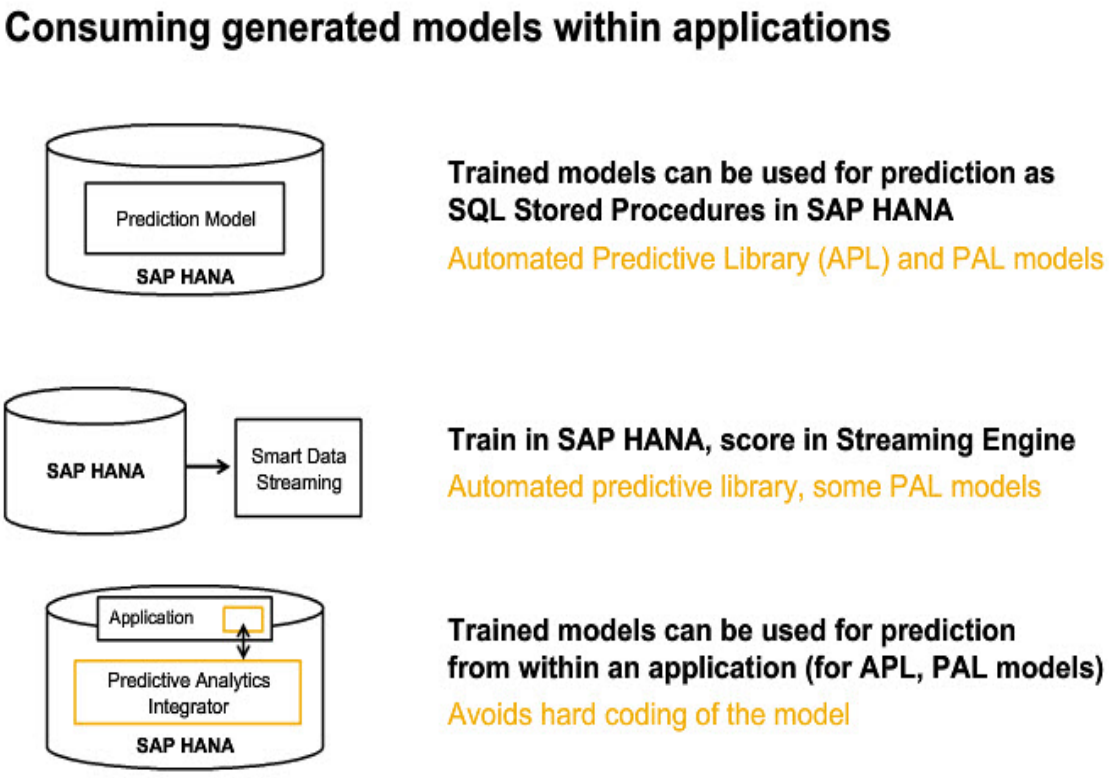
The choice between algorithmic (R and PAL) and automatic (APL) predictive capabilities for the most part depends on the target users and their wants. The APL provides flexibility to modify the predictive analytics progress while not users needing information of a way to build complicated information models from scratch. PAL or R usually needs a user to make procedures manually for every stage of the predictive modeling progress. information scientists naturally tend to like algorithmic Data scientists naturally tend to prefer algorithmic techniques that offer a high degree of control and precision in the modeling process. This flexibility comes at a cost: both R and PAL require users to be properly trained in data science techniques, as they must understand what each algorithm does, how it works, and how to interpret the results. Even seasoned data scientists less-sophisticated analyses have to be compelled to invest time to travel through the total predictive analytics advancement on every drawback once victimization recursive techniques. machine-controlled analytics automates several of the predictive-modeling steps that an information mortal generally performs for common workflows like classification, regression, and association analysis, saving the user time and energy. The machine-controlled machine learning engine still performs the total predictive analytics advancement however needs little or no input from the user. The result's a considerably quicker analysis that has fewer configuration parameters. In general, each information scientists and business analysts ought to begin their analysis by victimization the machine-controlled predictive capabilities of SAP HANA whenever potential machine-controlled machine learning will address a growing range of eventualities and generally will turn out valid leads to seconds or

minutes. this permits people who don't seem to be information scientists to answer their own queries and quickly retell on the leads to a self-service manner whereas giving information scientists an automatic method of analyzing several issues quickly. In some cases, wherever an information mortal might want to form an additional complicated model or be in complete management of every recursive parameter, it's applicable still to start out with APL. That way, the information mortal will perceive the information and make hypotheses before transitioning to Associate in Nursing recursive technique like SAP HANA–native PAL or off board R scripts.



Legend
APL = automated predictive library
PAL = predictive analysis library

Choosing SAP HANA as your predictive analytics platform provides you unmatched capabilities to perform native in-memory analytics with any combination of recursive (PAL) and automatic (APL) capabilities. each of those libraries run natively within the federation layer of SAP HANA and have direct access to the information. Calculations area unit performed at intervals SAP HANA, and so no information is extracted, no external I/O load is formed, and no external systems area unit needed. For cases during which an information mortal desires to use ASCII text file, third-party, or perhaps self-created R scripts, SAP HANA supports Associate in Nursing external. R server as a sidecar preparation. this permits much any R script to run on information from SAP HANA, however it needs you to extract the information, method it outwardly, and have the results sent back to the supply. Whenever potential, use technologies native to SAP HANA like the PAL and therefore the APL. Use Associate in Nursing External R Server solely as a final resort and for predictive calculations. Some information scientists could favor to do information manipulation in R yet, however this is able to be terribly inefficient compared with implementing those information manipulations at intervals a predictive tool or with a table or read from SAP HANA.



The prediction models are stored in SP HANA database and are used as SQL stored procedures in SAP HANA. The SAP HANA database using PAL library and some PAL models score it in streaming engine as shown in the figure. The trained models are then used for prediction from within another application by using APL, PAL models, this process avoids the hard coding of the model and saves time.

SAP predictive Analytics brings a replacement level of predictive analytics capabilities to organization. a large spectrum of users from business analysts to information scientists will quickly build automatic predictive models to realize on the-fly rating, perform time period predictive analysis, and infix models inside atomic number 83 workflows or the other business application. one platform allows everybody to profit from an automatic predictive resolution that encompasses the complete predictive lifecycle, from model creation to validation and readying, while not sacrificing speed, power, or quantifiably. With SAP predictive Analytics on the desktop, we have a tendency to gain a graphical desktop setting that supports all predictive technologies from SAP. It additionally ends the necessity to be told sophisticated SQL script cryptography or use developer-level tools to research information in SAP HANA. In most eventualities, a user doesn't need a mathematical or information science background to start out making predictive models quickly.

ORACLE (MACHINE LEARNIING)

The era of “big data” and therefore the “cloud” are driving corporations to vary. simply to stay pace, they need to learn new skills and implement new practices that leverage those new knowledge sources and technologies. Big data and analytics provide the promise to satisfy these new needs. Cloud, competition, Big data analytics and next-generation “predictive” applications are driving corporations towards achieving new goals of delivering improved “actionable insights” and higher outcomes. ancient atomic number 83 & Analytics approaches don’t deliver these elaborate prognosticative insights and easily can’t satisfy the rising client expectations during this new world order created by Big data and therefore the cloud.

Unfortunately, with Big data, because the knowledge grows and expands within the 3 V’s; rate, volume and selection (data types), new issues emerge. knowledge volumes grow and knowledge becomes unmanageable and unmovable. measurability, security, and knowledge latency become new problems. coping with unstructured knowledge, sensing element knowledge and spatial knowledge all introduce new knowledge kind complexities.

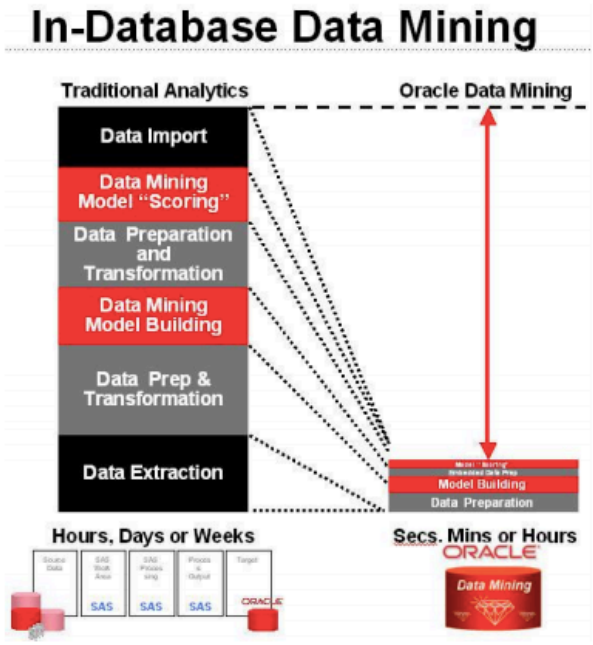
Traditional knowledge analysis usually starts with a sample or set of the information that's exported to separate analytical servers and tools (SAS, R, Python, SPSS, etc.) that are particularly designed for statisticians and knowledge scientists to research knowledge. The analytics they perform vary from easy descriptive applied mathematics analysis to advanced, prognosticative and prescriptive analytics. If a knowledge person builds a prognosticative model that's determined to be helpful and valuable, then IT must be concerned to work out readying and enterprise readying and application integration problems become subsequent huge challenge. The prognosticative model(s)—and all its associated knowledge preparation and transformation steps—have to be somehow translated to SQL and recreated within the knowledge the info the information base so as to use the models and create predictions on the larger datasets maintained within the data warehouse. This model translation section introduces tedious, time overwhelming and pricy manual secret writing steps from the initial applied mathematics language (SAS, R, and Python) into SQL. DBAs and IT should somehow “productionize” these separate applied mathematics models within the info and/or knowledge warehouse for distribution throughout the enterprise. Some vendors can charge for specialized merchandise and choices for only for prognosticative model readying. this can be wherever several advanced analytics comes fail. Add Hadoop, sensing element knowledge, tweets, and increasing Big data reservoirs and therefore the entire “data to unjust insights” method becomes more difficult.

Oracle Advanced Analytics offers a good library of powerful in-database algorithms and integration with open supply R that along will solve a good type of business issues and might be accessed via SQL, R or GUI. Oracle Advanced Analytics, Associate in Nursing choice to the Oracle info Enterprise Edition 12c, extends the info into Associate in Nursing enterprise-wide analytical platform for data-driven issues like churn prediction, client segmentation, fraud and anomaly detection, distinguishing cross-sell and upsell opportunities, market basket analysis, and text mining and sentiment analysis. Oracle Advanced Analytics empowers knowledge analyst, knowledge scientists and business analysts to additional extract data, discover new insights and create knowing predictions—working directly with giant knowledge volumes within the Oracle info. knowledge analysts/scientists have alternative and adaptability in however they move with Oracle Advanced Analytics. Oracle knowledge jack is Associate in Nursing Oracle SQL Developer extension designed for knowledge analysts that has a straightforward to use “drag and drop” advancement GUI to the Oracle Advanced Analytics SQL data processing functions (Oracle knowledge Mining). Oracle SQL Developer may be a free integrated development surroundings that simplifies the event and management of Oracle info in each ancient and Cloud deployments. once Oracle knowledge jack users ar happy with their analytical methodologies, they will share their workflows with alternative analysts and/or generate SQL scripts handy to their DBAs to accelerate model readying. Oracle knowledge jack additionally provides a PL/SQL API for advancement programming and automation. R programmers and knowledge scientists will use the acquainted open supply R applied mathematics artificial language console, R Studio or any IDE to figure directly with knowledge within the info and leverage Oracle Advanced Analytics’ R integration with the info (Oracle R Enterprise). Oracle Advanced Analytics’ Oracle R Enterprise provides clear SQL to R translation to equivalent SQL and Oracle data processing functions for in-database performance, correspondence, and scalability—this creating R prepared for the enterprise. Application developers, exploitation the ODM SQL data processing functions and ORE R integration will build fully machine-controlled prognosticative analytic solutions that leverage the strengths of the info and therefore the flexibly of R to integrate Oracle Advanced Analytics analytical solutions into atomic number 83 dashboards and enterprise applications. By integration Big data management and large knowledge analytics into identical powerful Oracle info knowledge

management platform, Oracle eliminates knowledge movement, reduces total price of possession and delivers the quickest thanks to deliver enterprise-wide prognostic analytics solutions and applications.

Oracle Advanced Analytics provides support for these knowledge driven issues by providing a large vary of powerful workhorse data processing algorithms that are enforced in an exceedingly electronic information service atmosphere (RDBMS). Algorithms square measure enforced as SQL functions within the info. Oracle Advanced Analytics' data processing algorithms thus leverage all connected SQL options and might mine knowledge in its original star schema illustration together with commonplace structured tables and views, transactional knowledge and aggregations, unstructured i.e. CLOB knowledge varieties (using Oracle Text to take apart out "tokens") and spacial knowledge. Oracle Advanced Analytics in-database SQL data processing functions make the most of correspondence within the info for each model build and model apply, honor all security and user privilege schemes, adhere to revision management and audit pursuit info options and might mine knowledge in its native and probably encrypted kind within the Oracle info.
Oracle data processing allows you to:

- Leverage your data to discover patterns and valuable new insights
- Build and apply predictive models and embed them into dashboards and applications
- Save money. Oracle Data Mining costs significantly less than traditional statistical software. A feature of the Oracle Database, ODM reduces the total cost of ownership.



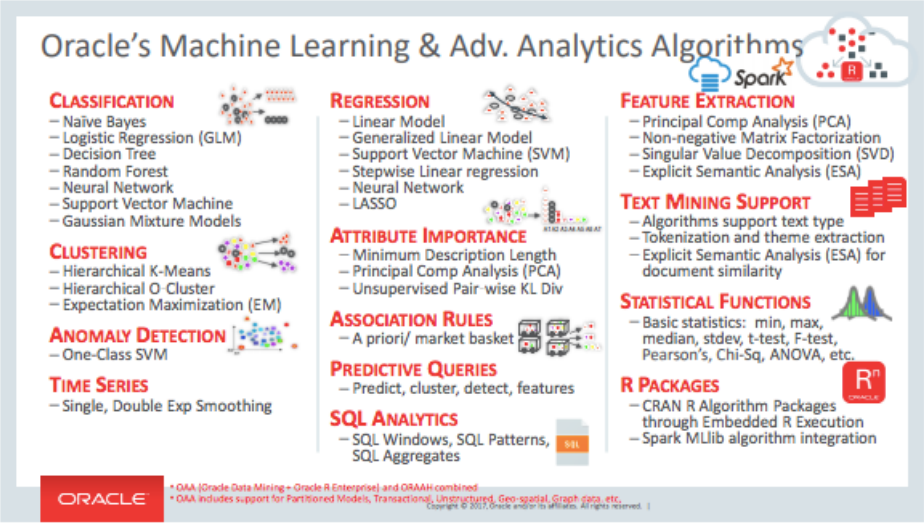
Oracle Data Mining eliminates data movement, data duplication and security exposures.

Predictive Analytics

Predictive analytics is the process of automatically sifting through large amounts of data to find previously hidden patterns, discover valuable new insights and make informed predictions for data-driven problems such as:

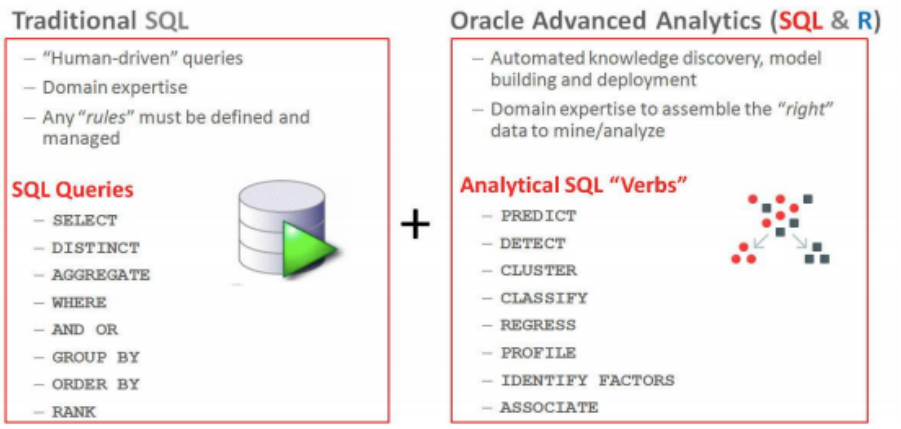
- Predicting customer behaviors, identifying cross-selling and up-selling opportunities
- Anticipating customer churn, employee attrition and student retention
- Detecting anomalies and combating potential tax, medical or expense fraud
- Understanding hidden customer segments and understanding customer sentiment, • Identifying key factors that drive outcomes and delivering improved quality

Oracle Advanced Analytics provides support for these knowledge driven issues by giving a large vary of powerful workhorse data processing algorithms that are enforced during a computer database atmosphere (RDBMS). Algorithms area unit enforced as SQL functions within the info. Oracle Advanced Analytics' data processing algorithms thence leverage all connected SQL options and may mine knowledge in its original star schema illustration together with customary structured tables and views, transactional knowledge and aggregations, unstructured i.e. CLOB knowledge varieties (using Oracle Text to dissect out "tokens") and spatial knowledge. Oracle Advanced Analytics in-database SQL data processing functions benefit of similarity within the info for each model build and model apply, honor all security and user privilege schemes, adhere to revision management and audit chase info options and may mine knowledge in its native and doubtless encrypted type within the Oracle info.



SQL & R Support

Most Oracle customers square measure terribly conversant in SQL as a language for question, reporting, and analysis of structured information. it's the factual customary for analysis and also the technology that underlies most atomic number 83 tools. R may be a wide fashionable open supply artificial language for applied math analysis that's free and since of that's educated in most information science academic programs. A growing range of knowledge analysts, information scientists, researchers, and teachers begin by learning to use R, resulting in a growing pool of R programmers UN agency will currently work with their information within the Oracle info victimization either SQL or R languages. Over the past decade and half, Oracle Advanced Analytics has matured and has been developed to currently in Oracle 12c, the Oracle Advanced Analytics possibility delivers nearly twenty ascendable, parallelized, in-database implementations of workhorse prognostic analytics algorithms. Oracle Advanced Analytics exposes these data processing algorithms as SQL functions that square measure accessible via SQL, R language and also the Oracle information laborer interface, associate degree extension to Oracle SQL Developer for the foremost common information driven issues e.g. clustering, regression, prediction, associations, text mining, associations analysis, etc. All Oracle Advanced Analytics algorithms square measure enforced deep within the info and take full advantage of the Oracle Database's business leading quantifiably, security, SQL functions, integration, ETL, Cloud, structured, unstructured and abstraction information sorts options and strengths and might be accessed via each SQL and R—and interface.



In Database Processing with Oracle Advanced Analytics

Oracle Advanced Analytics extends the info into a comprehensive advanced analytics platform for giant information analytics. With Oracle, powerful analytics square measure performed directly on information within the info. Results, insights, and real time prognostic models square measure accessible and managed by the info. a knowledge mining model may be a schema object within the info, engineered via a PL/SQL API that prepares the information, learns the hidden patterns to make associate degree OAA model which may then be scored via constitutional OAA data processing SQL functions. once building models, Oracle Advanced Analytics leverages existing ascendable technology (e.g., parallel execution, picture indexes, aggregation techniques) and extra developed new Oracle Advanced Analytics and Oracle info technologies (e.g., formula at intervals the parallel infrastructure, IEEE float, automatic information preparation for binning, handling missing values, support for unstructured information i.e. text, etc.).
The true power of embedding data processing functions at intervals the info as SQL functions is most evident once evaluation data processing models. Once the models are engineered by learning the hidden patterns within the historical information, applying the models to new information within the info is blazingly quick. evaluation is then simply a row-wise operate. Hence, Oracle Advanced Analytics will "score" several immeasurable records in seconds and is intended to support on-line transactional process (OLTP) environments.

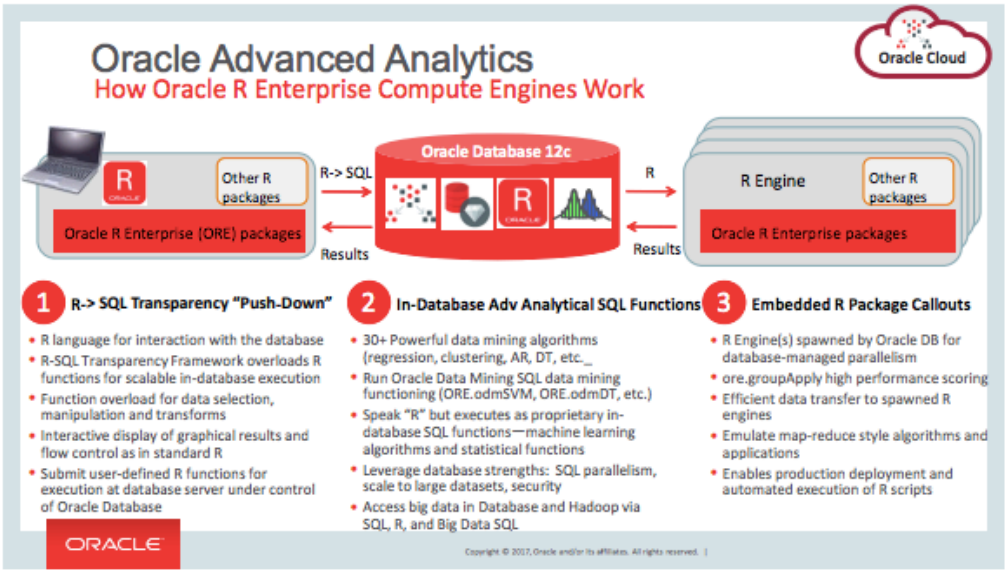
Typically, so as to perform correct analysis on knowledge, analysts ought to build specific choices concerning a way to "bin" knowledge, touch upon missing values and overtimes cut back the quantity of variables (feature selection) to be employed in the models. Over the past fifteen years, Oracle Advanced Analytics has evolved and currently will automatize most of the steps generally needed in data processing comes. Today, machine-driven knowledge Preparation (ADP) mechanically bins numeric attributes victimization default and user customizable binning ways e.g. equal breadth, equal count, user-defined and equally bins categorical attributes into N prime values and "other" or user-defined bins. Missing prices are mechanically replaced by an applied mathematics value (i.e. mean, median, mode, etc.) rather than that record being aloof from the analysis. ADP is employed each for model building and so once more for applying the models to new knowledge. Users will in fact override ADP settings if they select. Oracle Advanced Analytics provides support for attribute reduction (Attribute Importance victimization the Minimum Description Length algorithm) and have reduction techniques (Principal parts Analysis and Non-Negative Matrix Factorization). However, each of the Oracle Advanced Analytics algorithms (e.g. Decision Trees, Generalized Linear Regression, Support Vector Machines, Naïve Bayes, K-Means Clustering, Expectation Maximization Clustering, Anomaly Detection 1-Class SVMs, etc.) has their own built-in automated strategies for attribute reduction and selection so an explicit variable reduction step is optional, but not necessary. Users of course can control algorithm and data preparation settings or accept the intelligent defaults. Transactional data, e.g. purchases, transactions, events, etc. represent much of the data that is important to build good predictive models. Oracle Advanced Analytics mines this data in its native transactional form and leverages the database's aggregation functions to summarize it and then feed vector of the data (e.g. item purchases) and join it to other customer 2-D data to provide a 360-degree customer view. Oracle Advanced Analytics models, e.g. classification, regression and clustering models, ingest this aggregated transactional attribute as a "nested table". Deep inside the Oracle Advanced Analytics' in-database processing, records are processed as triplets: Unique ID, Attribute_name, and Attribute_value. That's just part of the secret sauce of how

Oracle Advanced Analytics leverages the core strengths of the Oracle Database. Market basket analysis would of course mine this data in its native transactional data form (typically not aggregated) to find co-occurring items in baskets. Unstructured data i.e. text is also processed in a similar fashion inside the database. Oracle Advanced Analytics uses Oracle Text's text processing capabilities and multi-language support to "tokenize" any CLOB data type e.g. text, Word, Adobe Acrobat, etc. As Oracle Text is a free feature in every Oracle Database, Oracle Advanced Analytics leverages it to pre-process unstructured data to then feed vectors of words and word coefficients (TFIDF— term frequency inverse document frequency) into the algorithms. Oracle Advanced Analytics just treats the unstructured attributes as additional input attributes e.g. police comments, physician's notes, resume, emails, article, abstract, etc. that get joined with everything else (e.g. Age, Income, Occupation, etc.) that is being fed into the Oracle Advanced Analytics data mining algorithms. Spatial data, web clicks and other data types can also be joined and included in Oracle Advanced Analytics data mining models.

Oracle R Enterprise - Integrating open source R with the Oracle Database

Oracle R Enterprise, a component of the Oracle Advanced Analytics leverages it to pre-process unstructured knowledge to then feed vectors of words and word coefficients (TFIDF— term frequency inverse document frequency) into the algorithms. Oracle Advanced Analytics simply treats the unstructured attributes as further input attributes e.g. police comments, physician's notes, resume, emails, article, abstract, etc. that get joined with everything else (e.g. Age, Income, Occupation, etc.) that's being fed into the Oracle Advanced Analytics data processing algorithms. spacial knowledge, net clicks and different knowledge varieties may be joined and enclosed in Oracle Advanced Analytics data processing models. Oracle R Enterprise -Integrating open supply R with the Oracle information

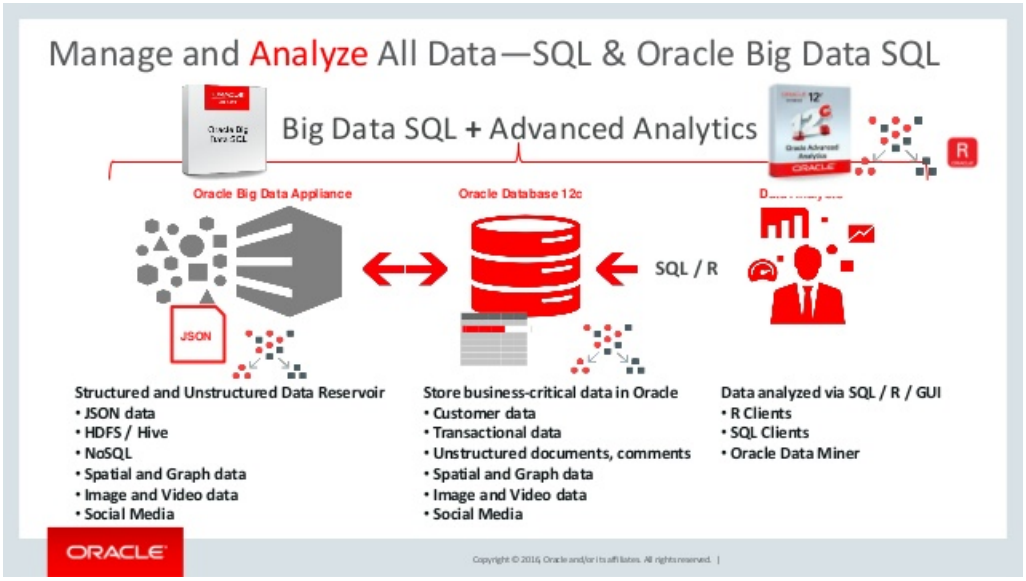
Oracle R Enterprise, an element of the Oracle Advanced Analytics possibility, makes the open supply R applied mathematics artificial language and surroundings prepared for the enterprise and large knowledge. "R provides a good sort of applied mathematics (linear and nonlinear modelling, classical applied mathematics tests, time-series analysis, classification, clustering,) and graphical techniques, and is extremely protractile. R's strengths area unit that it's free—open supply, powerful and protractile, has an in-depth array of graphical and applied mathematics packages and is continually being dilated by the R user community World Health Organization author and contribute R "packages". R's challenges area unit that it's memory strained, single rib, runs Associate in Nursing outer loop which will curtail process and isn't usually thought of to be "industrial strength". Contributed R packages area unit of variable quality. Oracle R Enterprise integrates R with Oracle information and maps R functions to equivalent SQL and Oracle data processing SQL functions and is intended for issues involving massive amounts of knowledge. it's a collection of R packages (ORE) Associate in Nursing Oracle information options that alter an R user to control on knowledgebase-resident data while not victimization SQL and to execute R scripts in one or a lot of embedded R engines that run on the information server. knowledge analysts and knowledge scientists will develop, refine, and deploy R scripts that leverage the correspondence and quantifiably of the information and also the SQL data processing functions to alter knowledge analysis in one step—without having to find out SQL Oracle R Enterprise has overladen open supply R ways and functions that transparently convert customary R syntax into SQL. These ways and functions area unit in ORE packages that implement the Oracle R Enterprise transparency layer. With these functions and ways, R programmers will produce R objects that access, analyze, and manipulate knowledge that resides within the information. The information mechanically optimizes the SQL code to boost the potency of the question. Oracle R Enterprise performs operate pushdown for in-database execution of base R, Oracle SQL applied mathematics functions, Oracle data processing SQL functions and elite in style R packages. as a result of it runs as Associate in Nursing embedded element of Oracle information, Oracle R Enterprise will run any R package either by operate pushdown or via "embedded R mode" whereas {the knowledge the info the information base manages the data served to the R engines. This "embedded R mode" ability permits developers to increase Oracle Advanced Analytics' natively supported toolkit with any open supply R packages and develop wide locomote and automatic advanced analytics methodologies that area unit utterly managed by the information.



Users, preferring to figure in R to access and analyze their knowledge, could use R Studio, or any R graphical user interface, to attach to Associate in Nursing Oracle information and access Oracle Advanced Analytics' R integration (Oracle R Enterprise). Once an association is created, the OAA/ORE session synchs the user's information in order that they see all their tables and views within the information. after they run any base R language operate it gets transparently mapped to equivalent SQL functions. R user's victimization the OAA/ODM algorithms and OAA/ORE algorithms will perform scalable data processing within the information.

Hadoop Oracle Big Data Appliance and Big Data SQL

Hadoop Oracle big data Appliance and large knowledge SQL big data is currently usually keep in Hadoop servers. The separate knowledge surroundings outside the information introduces new knowledge management and knowledge analysis challenges. Big data SQL addresses this challenge by extending SQL process to Hadoop via the Oracle Big data Appliance. victimization "smart scan" technology developed for Exadata, Big data SQL pushes down SQL logic to control on Hive tables. knowledge analysts will currently a lot of simply cash in on latest Big data sources of knowledge of knowledge of information of presumably unknown worth keep in Big data reservoirs and mix that knowledge with knowledge of best-known worth managed within an information and/or data warehouse. However, the data stored in Hadoop may be voluminous and sparse representation (transactional format) and lacking in information density. Given that much of the data may come from sensors, Internet of Things, "tweets" and other high-volume sources, users can leverage Big Data SQL to collect counts, maximum values, minimum values, thresholds count above or below user defined values, averages, shorter term averages and counts and longer time averages and counts, sliding SQL window averages and counts and comparisons of each to the other. So, filter "big data", reduce it, join it to other database data using Oracle Big Data SQL and then mine *everything* inside the Oracle Database using Oracle Advanced Analytics Option.



SQL and Big Data SQL enable data analysts to access, summarize, filter and aggregate data from both Hadoop servers and the Database and combine them for a more complete 360-degree customer view and build predictive models using Oracle Advanced Analytics.

Conclusion

Traditional BI and analytic approaches simply can't keep pace with requirements era of "big data" and "cloud". For organizations who strive to be leaders in their areas leveraging these new technologies, the prompt capture and collection of data of known and unknown value, the proper data management, assembly of relevant data and facile deep analysis and automation and deployment of the actionable insights is the key to success. Oracle Advanced Analytics, a priced option to the Oracle Database 12.2c, collapses the traditional extract, move, load, analyze, export, move, load/import paradigm all too common today. Oracle Advanced Analytics exposes these prognostic algorithms as SQL functions accessible via SQL (Oracle data processing OAA SQL API component), the Oracle knowledge jack "drag and drop" progress graphical user interface, associate degree extension to Oracle SQL Developer four.2 and thru tight integration w/ open supply R (Oracle R Enterprise R integration component). as a result of Oracle Advanced Analytics' in-database data processing machine learning/predictive analytics algorithms are engineered from the within out of the Oracle information and take full advantage of the Oracle Database's measurability, security, integration, cloud, structured and unstructured data processing capabilities, it makes Oracle the perfect platform for giant knowledge + analytics solutions and applications either on premise or on the Oracle Cloud.

With Oracle, knowledge management and descriptive, prognostic and prescriptive Big data analytics ae designed into the platform from the start. All of Oracle's multiple decades of vanguard knowledge management and SQL and massive knowledge SQL is harnesses and combined with Oracle's style and development approach of "moving the algorithms to the data" vs. "moving the info to the algorithms". Oracle's vision is to make an enormous knowledge and analytic platform for the time of Big data and therefore the cloud to:

Make big data + analytics simple:

- Any data size, on any computer infrastructure
- Any variety of data, in any combination

Make big data and analytics deployment simple:

- As a service, as a platform, as an application

By integrating both big data management and big data analytics into a single unified Oracle Database platform, Oracle reduces total cost of ownership, eliminates data movement, and delivers the fastest way to deliver enterprise-wide predictive analytics solutions and applications.

Future of Machine Learning

Despite its sensible and industrial successes, machine learning remains a young field with several underexplored analysis opportunities. a number of these opportunities may be seen by contrastive current machine-learning approaches to the kinds of learning we tend to observe in present systems like humans and alternative animals, organizations, economies, and biological evolution. Some researchers to start exploring the question of the way to construct laptop long or endless learners that operate nonstop for years, learning thousands of interconnected ability's or functions at intervals associate degree overall design that enables the system to boost its ability to be told one skill supported having learned. Another side of the analogy to natural learning systems suggests the thought of team-based, mixed-initiative learning. as an example, whereas current machine learning systems generally operate in isolation to research the given knowledge, folks usually add groups to gather and analyze knowledge (e.g., biologists have worked as groups to gather and analyze genomic knowledge, transfer along numerous experiments and views to form progress on this tough problem). New machine-learning strategies capable of operating collaboratively with humans to together analyze advanced knowledge sets may assemble the skills of machines to tease out delicate applied math regularities from big data sets with the skills of humans to draw on numerous background to come up with plausible explanations and counsel new hypotheses. like any powerful technology, machine learning raises questions about that of its potential uses society ought to encourage and discourage. The push in recent years to collect new kinds of personal data, motivated by its economic value, leads to obvious privacy issues, as mentioned above. The increasing value of data also raises a second ethical issue: Who will have access to, and ownership of, online data, and who will reap its benefits? Currently, much data is collected by corporations for specific uses leading to improved profits, with little or no motive for data sharing. However, the potential benefits that society could realize, even from existing online data, would be considerable if those data were to be made available for public good. Machine learning is likely to be one of the most transformative technologies of the 21st century. Although it is impossible to predict the future, it appears essential that society begin now to consider how to maximize its benefits.

References

1. https://www.thenextview.nl/blog/machinelearning
2. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/
3. https://www.thenextview.nl/blog/machine-learning-with-sap
4. https://blogs.sap.com/2017/07/18/machine-learning-in-sap-hana-asug-webcast-summary/
5. http://www.nordicdatasciencesummit.com/assets/whitepappers/the_powerofpredictivetext.pdf
6. http://www.oracle.com/technetwork/database/options/advanced-analytics/oaa122whitepaperv2-3787080.pdf
7. Images Courtesy of Google

Work Done By:

- **Anupam Sahay** (CIN :305903512)

  INTRODUCTION TO MACHINE LEARNING AND ALGORITHMS *(Page1 to Page 9)*

- **Krithy Nanaiah Atrangada** (CIN:305903525)

  SAP HANA MACHINE LEARNING *(Page 10 – Page 17)*

- **Tejas Agara Chandrakumar** (CIN: 306594462)

  ORACLE MACHINE LEARNING *(Page 18 – Page24)*