# The basics of Machine Learning

Heli Helskyaho

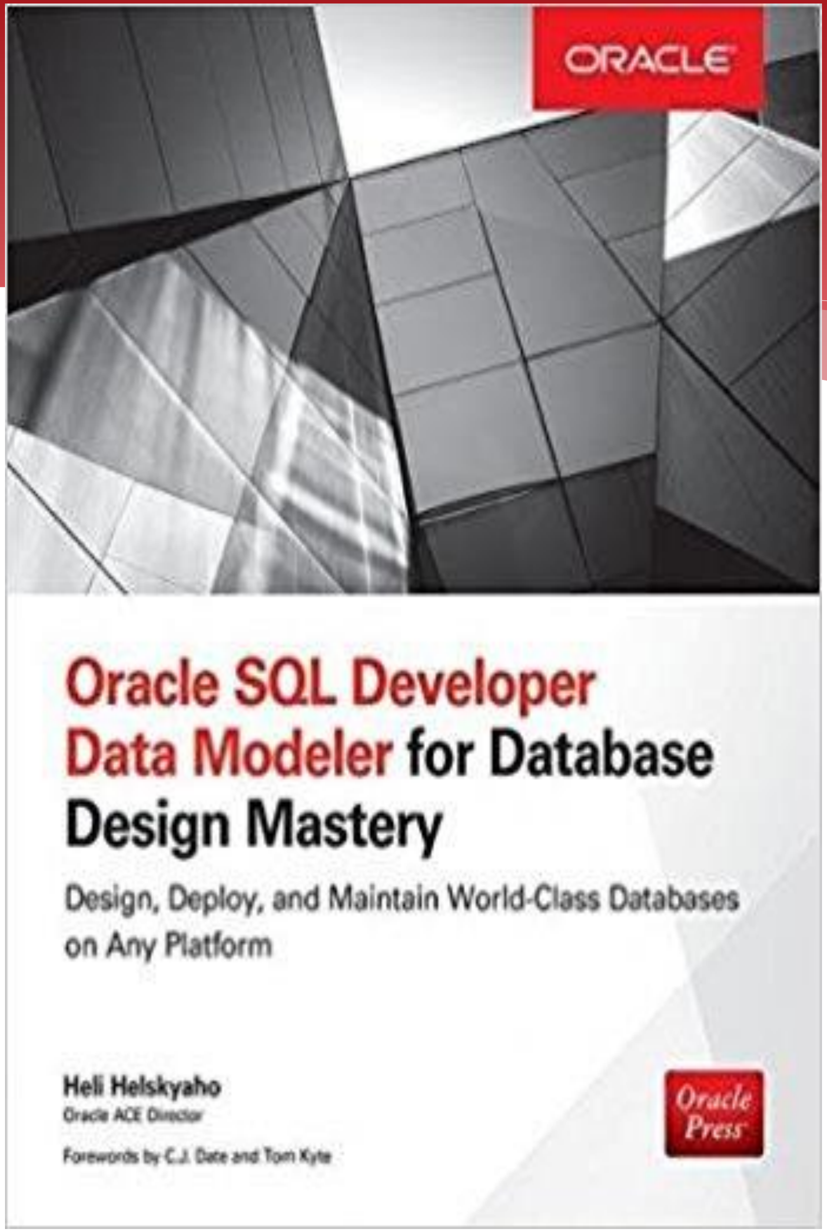
Nordic ACE Tour 2017

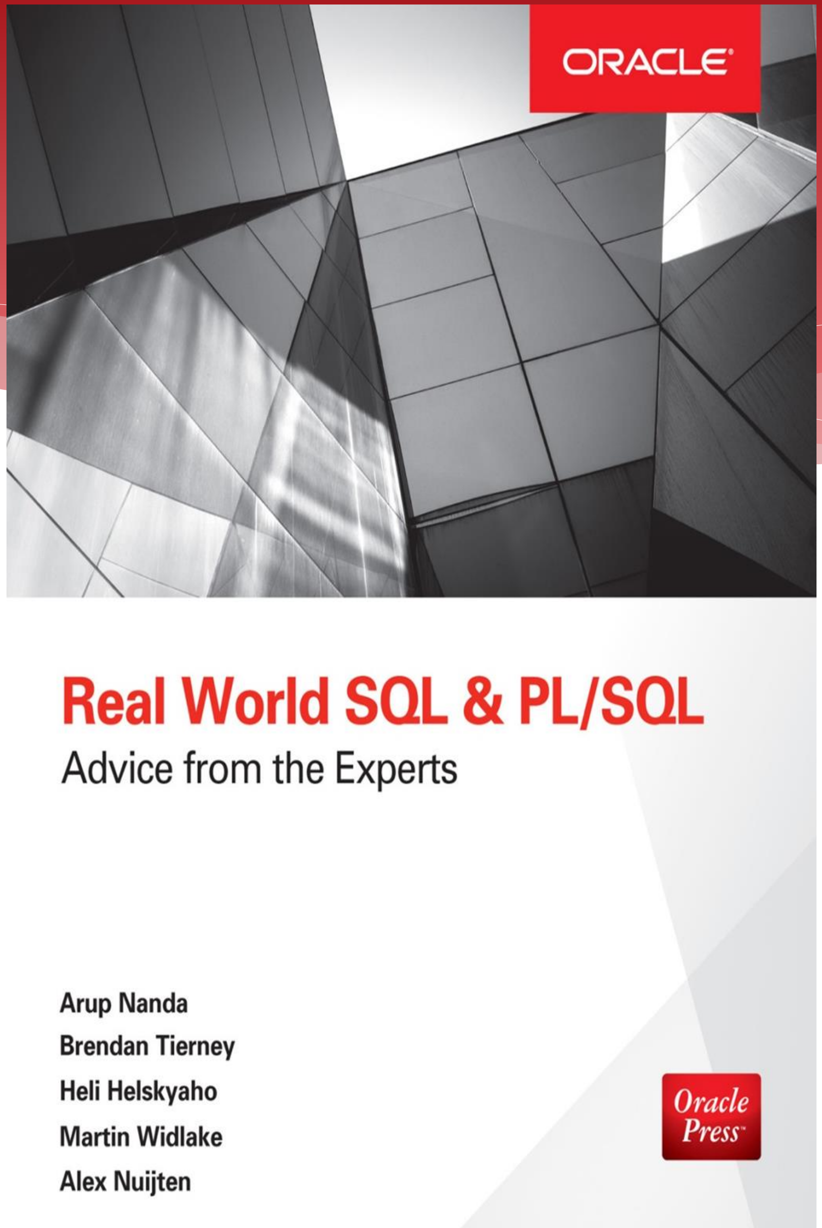
Miracle Finland Oy

# Introduction, Heli

* Graduated from University of Helsinki (Master of Science, computer science), currently a doctoral student, researcher and lecturer (databases, Big Data, Multi-model Databases, methods and tools for utilizing semi-structured data for decision making) at University of Helsinki
* Worked with Oracle products since 1993, worked for IT since 1990
* Data and Database!
* CEO for Miracle Finland Oy
* Oracle ACE Director
* Ambassador for EOUC (EMEA Oracle Users Group Community)
* Public speaker and an author
* Winner of Devvy for Database Design Category, 2015
* Author of the  book Oracle SQL Developer Data Modeler for Database Design Mastery (Oracle Press, 2015), co-author for Real World SQL and PL/SQL: Advice from the Experts (Oracle Press, 2016)

Miracle Finland Oy




MIRACLE
Miracle Finland Oy

# 500+ Technical Experts Helping Peers Globally

ORACLE ACE PROGRAM

ORACLE ACE Director

ORACLE ACE

ORACLE ACE Associate

**3 Membership Tiers**

- Oracle ACE Director
- Oracle ACE
- Oracle ACE Associate

bit.ly/OracleACEProgram

**Connect:**

oracle-ace_ww@oracle.com

Facebook.com/oracleaces

@oracleace

ORACLE Developer Community

Nominate yourself or someone you know: **acenomination.oracle.com**

# What is Machine Learning?

* An important part of Artificial Intelligence (AI)
* Machine learning (ML) teaches *computers* to learn from *experience (algorithms)*
  * Learn from data and make predictions
  * Mathematics, statistics,…
* “field of study that gives computers the ability to learn without being explicitly programmed“

-- Arthur Samuel, 1959

* A systematic study of algorithms and systems that improve their *knowledge* or *performance* with *experience*

MiRACLE
Miracle Finland Oy

# Why ML? Why now?

* Improved technology
* The price for storage solutions
* …
* An environment that NEEDS ML and is finally able to really use it

* Artificial Intelligence (AI)
* BIG DATA

Miracle Finland Oy

# What is Big Data?

* There is *no size* that makes a data to be ”Big Data”, it always depends on the capabilities
* The data is ***”Big”*** when traditional processing with traditional tools is not possible due to the amount or the complexity of the data
  * You cannot open an attachement in email
  * You cannot edit a photo
  * etc.

Miracle Finland Oy

# The three V's

* **Volume**, the size/scale of the data
* **Velocity**, the speed of change, analysis of streaming data
* **Variety**, different formats of data sources, different forms of data; structured, semi-structured, unstructured

# The other V's

* **Veracity**, the uncertainty of the data, the data is worthless or harmful if it's not accurate
* **Viability**, validate that hypothesis before taking further action (and, in the process of determining the viability of a variable, we can expand our view to determine other variables)
* **Value**, the potential value
* **Variability,** refers to data whose meaning is constantly changing, in consistency of data; for example words and context
* **Visualization**, a way of presenting the data in a manner that's readable and accessible

Miracle Finland Oy

# Challenges in Big Data

* More and more data (volume)
* Different data models and formats (variety)
* Loading in progress while data exploration going on (velocity)
* Not all data is reliable (veracity)
* We do not know what we are looking for (value, viability, variability)
* Must support also non-technical users (journalists, investors, politicians,… ) (visualization)
* All must be done *efficiently and fast and as much as possibly by machines*

Miracle Finland Oy

# When to use ML?

* You have **data**!
  * ML cannot be performed without data
  * part of the data for finding the model, part to prove it (not all for finding the model!)
* Rules and equations are
  * Complex (image recognition)
  * Constantly changing (fraud detection)
* The nature of the data changes and the program must adapt (today's spam is tomorrow's ham) (predicting shopping trends)

Miracle
Miracle Finland Oy

# Real life use cases for ML

* Spam filters
* Log filters (and alarms)
* Data analytics
* Image recognition
* Speech recognition
* Medical diagnosis
* Robotics
* …

MIRACLE Miracle Finland Oy

# Approximation!

* ML always gives an approximated answer
* Some are better than others, some are useful

* search for patterns and trends
* Prediction accuracy: the higher the number the better it will work on new data
* several models, choose the best, but still: all approximations! There is no correct answer…

Miracle Finland Oy

# What do I find the most difficult for a beginner?

* The terms!
  * So many different terms
  * The same term meaning different things, two (or more) terms for the same thing (sometimes a completely different word, sometimes just a short of the original word)
  * The relationships the terms have

# Terms used 1/5

* A Task
  * The problem to be solved with ML
* An Algorithm
  * the “experience” for the computer to learn with, solves the learning problem
  * Produces the Models

Miracle Finland Oy

# Terms used 2/5

* A Model
  * The output of ML
  * The Task is Addressed by Models

Miracle Finland Oy

# Terms used 3/5

* Different Models:
  * Predictive model
    * the model output involves the target variable
    * ” forecast what might happen in the future”
  * Descriptive model
    * the model output does not involve the target variable
    * ”what happened”
  * Prescriptive model
    * recommending one or more courses of action and showing the likely outcome of each decision
    * A predictive model + actionable data and a feedback system to track the outcome

# Terms used 4/5

* Different models based on the algorithm type:
  * Classification Models
  * Concept learning Models
  * Tree Models
  * Rule Models
  * Linear Models
  * Distance-based Models
  * Probabilistic Models

# Terms used 5/5

* Features/Dimensions
  * an individual *measurable property* or *characteristic of a phenomenon* being observed (Bishop, Christopher (2006), Pattern recognition and machine learning)
  * *Deriving features* (feature engineering, feature extraction) is one of the most important parts of machine learning. It turns data into information that a machine learning algorithm can use.
* Methods/Techniques
  * Unsupervised learning
  * Supervised learning

MIRACLE
Miracle Finland Oy

# The Task

* It is very important to define the Task well
* Machine learning is not only a computational subject, the practical side is very important

# It’s all about Algorithms

* Humans learn with *experience*, machines learn with *algorithms*
* It is not easy to find the right Algorithm for the Task
  * usually try with several algorithms to find the best one
  * selecting an algorithm is a process of trial and error

# Which algorithm?

* The selection of an algorithm depends on for instance
  * the size and type of data
  * the insights you want to get from the data
  * how those insights will be used
* It’s a trade-off between many things
  * Predictive accuracy on new data
  * Speed of training
  * Memory usage
  * Transparency (black box vs “clear-box”, how decisions are made)
  * Interpretability (the ability of a human to understand the model)
  * …

# Models 1/2

* Geometric models
  * Support vector machines, SVM
  * Notion of distance: Euclidean distance, nearest-neighbour classifier, Manhattan distance
* Probabilistic models
  * Bayesian classifier
* Logical models
  * Decision trees

# Models 2/2

* Grouping models, number of groups determined at the training time
  * Tree based models
* Grading models, ”infinite” resolution
  * Geometric classifiers
* …

# Features

* A Model is only as good as its Features…
* Interaction between features

* The unnecessary detail can be removed by discretisation (11,1kg vs 10kg)

# ML in short

* Use the right *Features*
  * with right Algorithms
    * to build the right *Models*
      * that archive the right *Tasks*

# Two types of Methods

* Unsupervised learning
  * finds hidden patterns or intrinsic structures in input data
* Supervised learning
  * trains a model on known input and output data to predict future outputs

# Unsupervised Learning

* Learning from unlabeled input data by finding hidden patterns or intrinsic structures in that data
* Machine learning algorithms find natural patterns in data to make better decisions and predictions possible
* used typically when you
  * don’t have a specific goal
  * are not sure what information the data contains
  * want to reduce the features of your data as a preprocessing for supervised learning

# Clustering

* *Clustering* is the most common method for unsupervised learning and used for *exploratory data analysis to find hidden patterns or groupings in data.*
* *Clustering algorithms*
  * *Hard clustering*
    * each data point belongs to *only one* cluster
  * *Soft clustering*
    * each data point can belong to *more than one* cluster

# Hard clustering algorithms

* each data point belongs to *only one* cluster

# Some Hard Clustering Algorithms 1/2

* K-Means (Lloyd's algorithm)
  * Partitions data into k number of mutually exclusive clusters (centroids)
  * Assign each observation to the closest cluster
  * Move the centroids to the true mean of its observations
  * When to use:
    * When the number of clusters is known
    * Fast clustering of large data sets
* K-Medoids
  * Similar to k-means, but with the requirement that the cluster centers coincide with points in the data (chooses datapoints as centers, medoids).
  * Can be more robust to noise and outliers than K-Means
  * When to use:
    * When the number of clusters is known
    * Fast clustering of categorical data

MIRACLE Miracle Finland Oy

# Some Hard Clustering, Algorithms 2/2

* Hierarchical Clustering
  * Divisive method, assign all observation to one cluster and the partition that cluster
  * Agglomerative method, each observation to its own cluster and merge those clusters
  * When to use:
    * When you don’t know in advance how many clusters
    * You want visualization to guide your selection

# Soft clustering algorithms

* each data point can belong to *more than one* cluster

Miracle Finland Oy

# Some Soft clustering algorithms

* Fuzzy C-Means (FCM)
  * Similar to k-means, but data points may belong to more than one cluster.
  * When to use:
    * The number of clusters is known
    * When clusters overlap
    * Typically for pattern recognition
* Gaussian Mixture Model
  * Partition-based clustering where data points come from different multivariate normal distributions with certain probabilities. (example: Prices for a house in different area)
  * When to use:
    * Data point might belong to more than one cluster
    * Clusters have different sizes and correlation structures within them

Miracle Finland Oy

# Supervised Learning

* Learning from known, labelled data
* Training a model on known input and output data to predict future outputs (remember that uncertainty is always involved)

# A process of supervised learning 1/2

1. Train
   1. Load data
   2. Pre-process data
   3. Learn using a method and an algorithm
   4. Create a model
   * iterate until you find the best model

# A process of supervised learning 2/2

2. Predict (use the model with new data)
   1. New data
   2. Pre-process data
   3. Use the model
   4. Get predictions
   5. Integrate the models into applications

# Supervised Learning, methods/techniques

* Predictive models
  * Classification
  * Regression

# Supervised Learning, Classification

* Classification models are trained to *classify* data into *categories*.
* They predict discrete responses
  * an email is genuine or spam
  * a tumor is small, medium size, or large
  * a tumor is cancerous or benign
  * a person is creditworthy or not
* For example applications like medical imaging, speech recognition, and credit scoring

# Supervised Learning, Classification

* Can the data be tagged or categorized? Can it be separated into specific groups or classes?
  * Classification might be the right answer
* Is the problem binary or multiclass?
  * Defines the number of classes.

MiRACLE Miracle Finland Oy

# Classification,
# Some Algorithms

* k Nearest Neighbor (kNN)
  * kNN categorizes objects based on the classes of their nearest neighbors all ready categorized
  * kNN predictions assume that objects near each other are similar
  * When to use:
    * need a simple algorithm to establish benchmark learning rules
    * memory usage of the trained model is a lesser concern (can be very memory consuming)
    * prediction speed of the trained model is a lesser concern (can be slow if the amount of data is large or several dimensions are used)

Miracle Finland Oy

# Classification, Some Algorithms

- Naïve Bayes
  - assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature when the class is defined
  - classifies new data based on the highest probability of its belonging to a particular class (a fruit is red -> an apple, a fruit is round -> an apple, together a stronger probability to be an apple)
  - When to use:
    - For a dataset containing many parameters (dimensionality of the inputs is high)
    - Simple to implement, easy to interpret

# Classification,
# Some Algorithms

* Discriminant Analysis
  * The classes are known a prio, an observation is classified to into one class based on the measured characteristics.
    * Example, bank notes:
    * two populations of bank notes, genuine, and counterfeit
    * Six measures:  length, right-hand width,  left-hand width,  top margin,  bottom margin, diagonal across the printed area
    * Take a bank note of unknown origin and determine using these six measurements whether or not it is real or counterfeit.
  * When to use:
    * need a simple model that is easy to interpret
    * memory usage during training is a concern
    * need a model that is fast to predict

MiRACLE
Miracle Finland Oy

# Classification, Some Algorithms

* Neural Network
  * Imitates how biological nervous systems, the brain, process information
  * A large number of highly interconnected processing elements (neurones) work together to solve specific problems
  * When to use:
    * For modeling highly nonlinear systems
    * When data is available incrementally and you wish to constantly update the model
    * Unexpected changes in your input data may occur
    * Model interpretability is not a key concern

Miracle Finland Oy

# Classification, Some Algorithms

* Decision Trees, Bagged and Boosted Decision Trees
  * A tree consists of branching conditions, predict responses to data by following the decisions in the tree from the root down to a leaf node
  * A bagged decision tree consists of several trees that are trained independently on data. Boosting involves reweighting of misclassified events and building a new tree with reweighted events.
  * When to use:
    * Need an algorithm that is easy to interpret and fast to fit
    * To minimize memory usage
    * High predictive accuracy is not a requirement
    * The time taken to train a model is less of a concern

Miracle Finland Oy

# Classification, Some Algorithms

* Logistic Regression
  * Predict the probability of a binary response belonging to one class or the other
    * For example how does hours spent studing affect the probability for  a student to pass the exam (yes/no)
  * When to use:
    * When data can be clearly separated by a single, linear boundary
    * Logistic regression is commonly used as a starting point for binary classification problems
    * As a baseline for evaluating more complex classification methods

Miracle Finland Oy

# Supervised Learning, Regression

* To predict continuous responses
  * changes in temperature
  * fluctuations in electricity demand
* For example applications like forecasting stock prices, handwriting recognition, acoustic signal processing, failure prediction in hardware, and electricity load forecasting.

# Regression, Some Algorithms

* Linear Regression
  * used to describe a continuous response variable as a linear function of one or more predictor variables
  * When to use:
    * need an algorithm that is easy to interpret and fast to fit, often the first model to be fitted to a new dataset
    * As a baseline for evaluating other, more complex, regression models

# Regression, Some Algorithms

* Nonlinear Regression
  * describe nonlinear relationships in experimental data

  * When to use:
    * When data has nonlinear trends and cannot be easily transformed into a linear space
    * For fitting custom models to data

MIRACLE
Miracle Finland Oy

# Regression,
# Some Algorithms

* Generalized Linear Model (GLM)
  * A special case of nonlinear models that uses linear methods: it fits a linear combination of the inputs to a nonlinear function (the link function) of the outputs
  * When to use:
    * When the response variables have non-normal distributions

# Regression, Some Algorithms

* Gaussian Process Regression Model (GPR)
  * nonparametric models that are used for predicting the value of a continuous response variable
  * When to use:
    * For interpolating spatial data
    * As a surrogate model to facilitate optimization of complex designs such as automotive engines
    * Can be used for example forecasting of mortality rates

# Regression, Some Algorithms

* Regression Tree
  * Decision trees for regression are similar to decision trees for classification, but they are modified to be able to predict continuous responses
  * When to use:
    * When predictors are categorical (discrete) or behave nonlinearly

# Improving Models

* Why to improve
  * To increase the accuracy and predictive power of the model
  * To increase the ability to recognize data from noise
  * To increase the performance
  * To improve the Measures wanted
  * …

# Improving Models

* Model improvement involves
  * Feature engineering
    * Feature selection
    * Feature transformation/extraction
  * Hyperparameter tuning

# Feature selection

* Also called variable selection or attribute selection
  * Identifying the most relevant features that provide the best predictive model for the data
  * *Adding* variables to the model to improve the accuracy or *removing* variables that do not improve model performance

# Feature selection techniques

* **Stepwise regression:**
  * adding or removing features sequentially until there is no improvement in prediction accuracy
* **Sequential feature selection:**
  * adding or removing predictor variables iteratively and evaluating the effect of each change on the performance of the model
* **Regularization:**
  * Using shrinkage estimators to remove redundant features by reducing their weights (coefficients) to zero
* **Neighborhood component analysis (NCA):**
  * Finding the weight each feature has in predicting the output, so that features with lower weights can be discarded

# Feature transformation

* Feature transformation is a form of *dimensionality reduction*
* Used when
  * want to reduce the dimensions/features of your data as a preprocessing for supervised learning
  * As datasets get bigger, you frequently need to reduce the number of features, or dimensionality.

# Feature transformation

* Techniques:
  * Principal component analysis (PCA)
  * Factor analysis
  * Non-negative matrix factorization

# Principal component analysis (PCA)

* Converts a set of observations of possibly correlated variables into a smaller set of values of linearly uncorrelated variables called *principal components*
* The first principal component will capture the most variance, followed by the second principal component, and so on.

# Factor analysis

* identifies underlying correlations between variables in a dataset to provide a representation in terms of a smaller number of unobserved variables, factors

# Non-negative matrix factorization (NNMF)

* Also called non-negative matrix approximation
* used when model elements must represent *non-negative* quantities, such as physical quantities

# Hyperparameter tuning

* Also called as Hyperparameter optimization
* Choosing an optimal set of hyperparameters for a learning algorithm
  * Hyperparameters are parameters whose values are set *prior* to the commencement of the learning process (the value of other parameters is derived via training)
  * Hyperparameters control how a machine learning algorithm fits the model to the data.

MIRACLE
Miracle Finland Oy

# Hyperparameter Tuning

* Tuning is an iterative process
  * Set parameters based on a best guess
  * Aim to find the best possible values to yield the best model
  * As you adjust hyperparameters and the performance of the model begins to improve, you see which settings are effective and which still require tuning
* Some examples of optimization algorithms:
  * Grid search
  * Bayesian optimization
  * Gradient-based optimization
  * Random Search
* A simple algorithm with well-tuned parameters is often better than an inadequately tuned complex algorithm, in many ways.

MIRACLE Miracle Finland Oy

# How do I know when to tune?

* How does the model perform on the data?
* Which of the models is the best?
* Which of the learning algorithms gives the best model for the data?
* …
* To be able to answer questions like these we need to have **measuring**

Miracle Finland Oy

# What to measure?

* Number of positives, number of negatives, number of true positives, number of false positives, number of true negatives, number of false negatives
* Portion of positives, portion of negatives
* Class ratio
* Accuracy, Error rate
* ROC curve, coverage curve,
* …
* It all depends on how we define the performance for the answer to our question (experiment): *experimental objective*

Miracle Finland Oy

# How to measure?

* And how to interpret?
* It all depends what we are measuring…
* Example: Testing the model accuracy
  * Tool: Cross validation

# Cross validation

* Sometimes called Rotation Estimation
* Divide the data in n parts of equal size
* Use n-1 parts for training and 1 for testing
* Repeat n times so that each of the sets will be used for testing

# What's next to learn?

* There is still so much more about ML…
* Reinforcement learning
  * the machine or software agent learns based on feedback from the environment
* Preference learning
  * inducing predictive preference models from empirical data
* Multi-task learning
  * multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks
* Online machine learning
  * data becomes available in a sequential order and is used to update our best predictor for future data at each step

MIRACLE
Miracle Finland Oy

# What's next to learn?

* Active learning
  * A learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points
* Deep learning
  * Images and anything that is in ”several layers”
* Adaptive Intelligence
  * People and machines

Miracle Finland Oy

# Data Mining

* building machine learning models is an essential step in the data mining process

Miracle Finland Oy

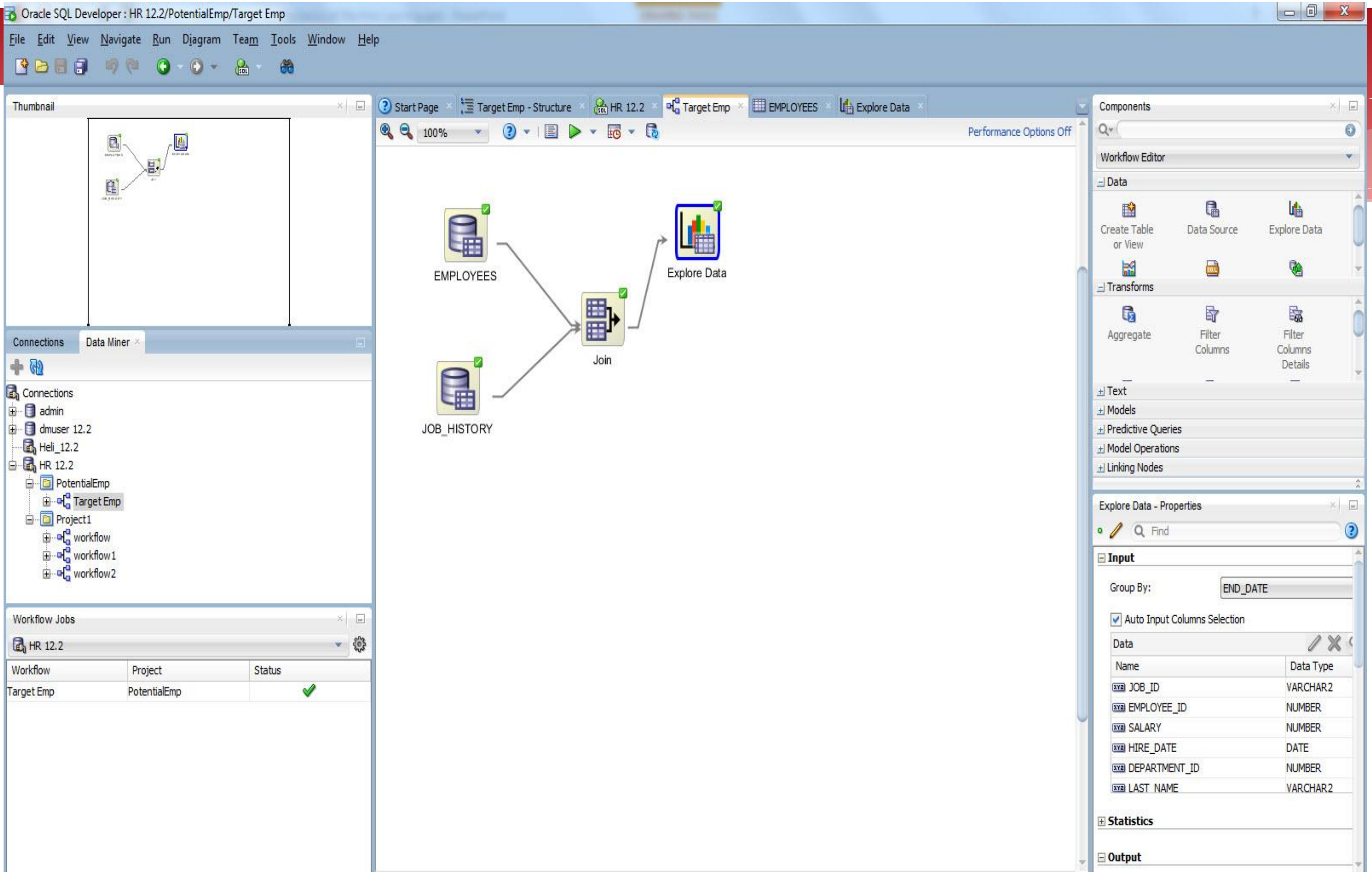
# Oracle SQL Developer, Data Miner

* Oracle SQL Developer is a free tool from Oracle
* Has an add-on called Data Miner
* Oracle Data Miner GUI Installation Instructions

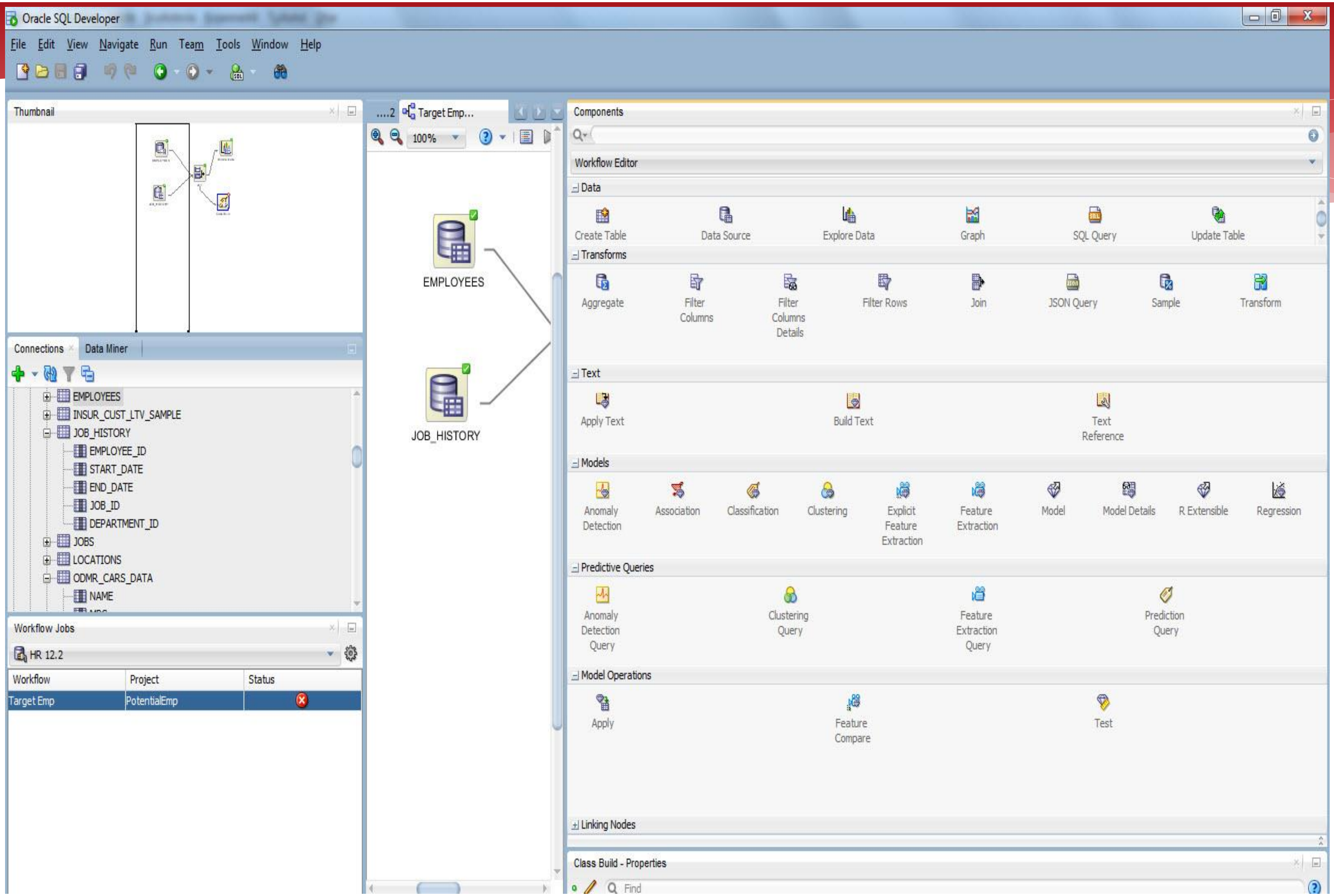
http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odmrinstallation-2080768.html

* Tutorial

http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/12c/BigDataDM/ODM12c-BDL4.html

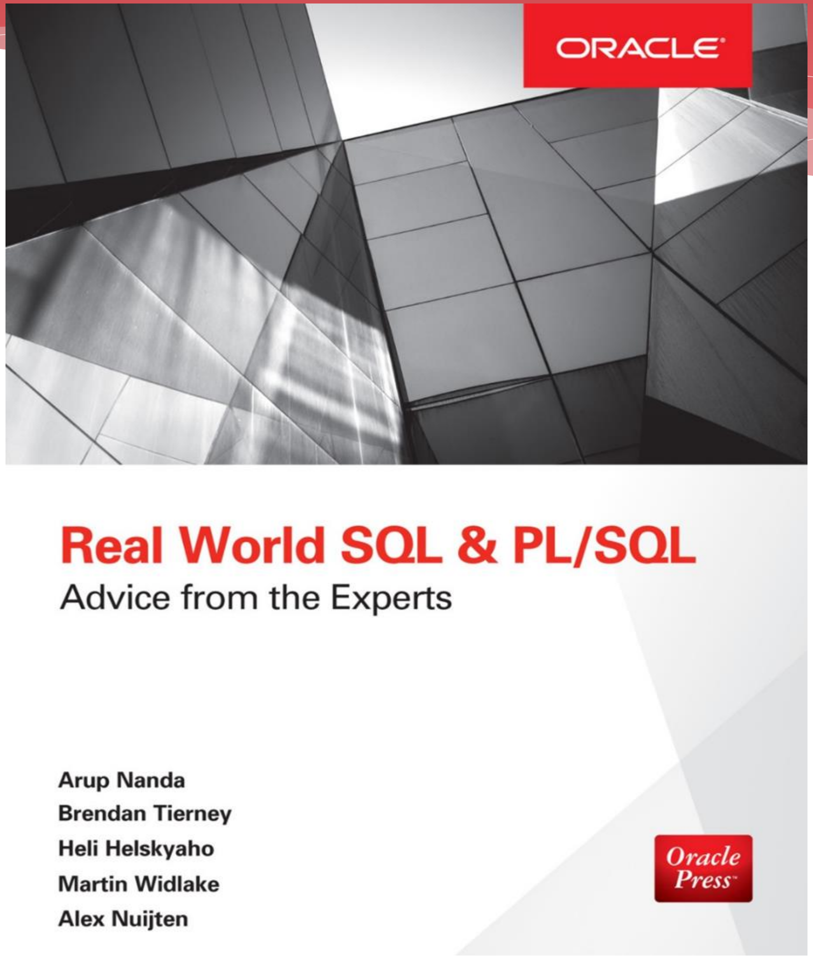



MIRACLE
Miracle Finland Oy

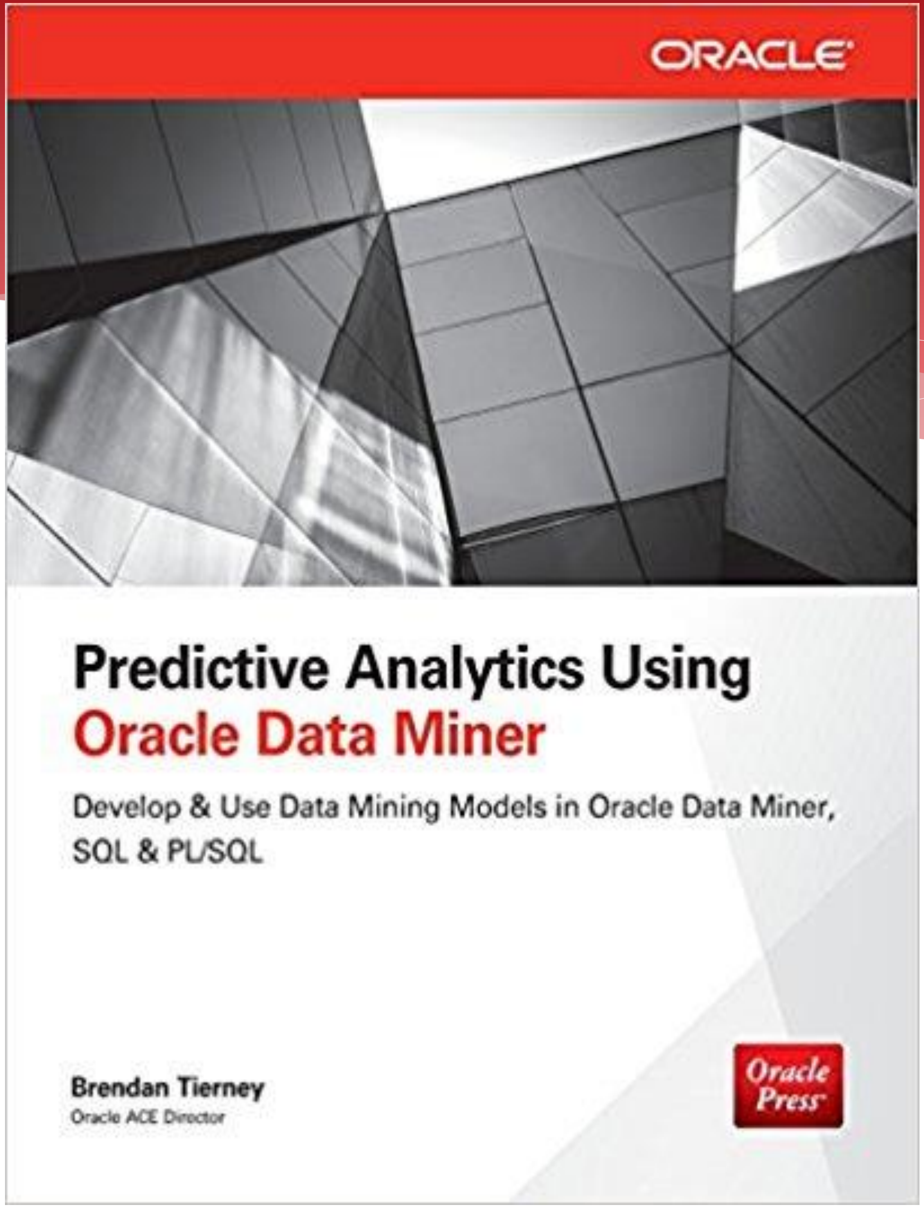



MIRACLE
Miracle Finland Oy

# Chapter 10




MIRACLE
Miracle Finland Oy

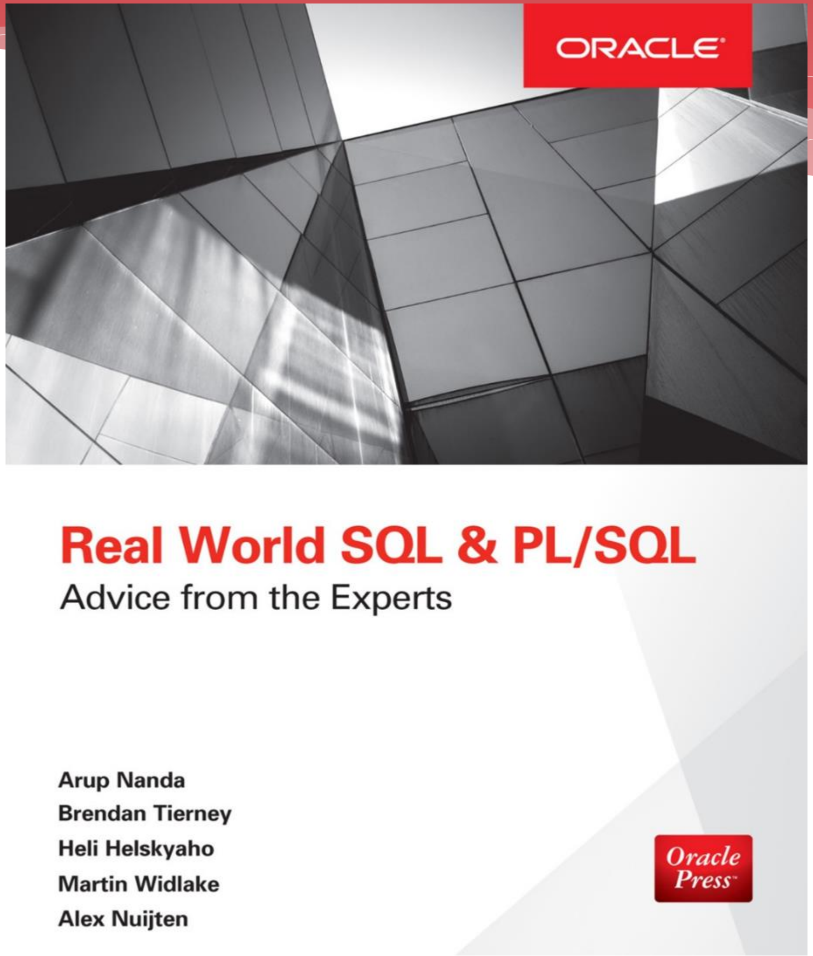

ORACLE
Predictive Analytics Using
Oracle Data Miner
Develop & Use Data Mining Models in Oracle Data Miner, SQL & PL/SQL
Brendan Tierney
Oracle ACE Director
Oracle Press

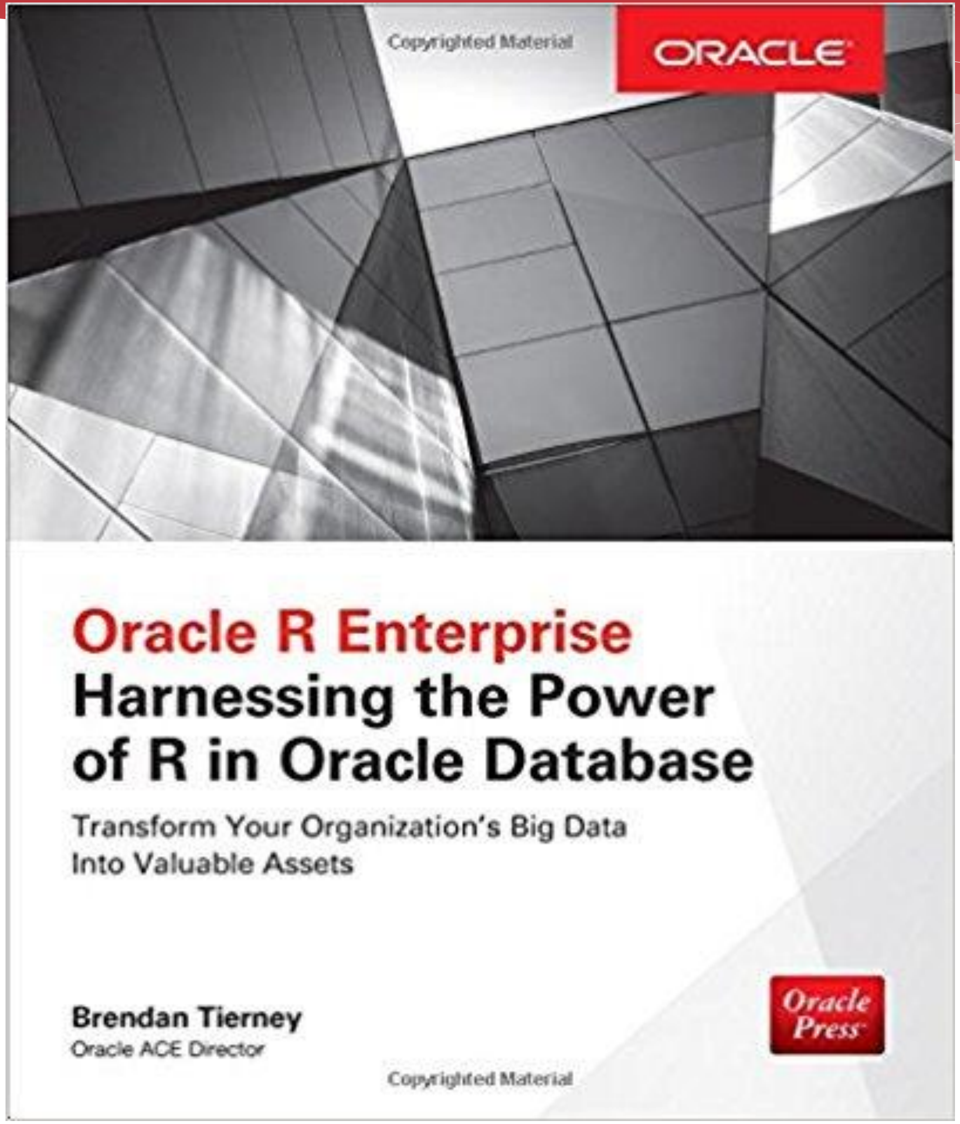
# Oracle R Enterprice

* a component of the Oracle Advanced Analytics Option (payable option)
* open source R statistical programming language in an Oracle database

MIRACLE
Miracle Finland Oy

# Chapter 11


ORACLE
Real World SQL & PL/SQL
Advice from the Experts
Arup Nanda
Brendan Tierney
Heli Helskyaho
Martin Widlake
Alex Nuijten
Oracle Press

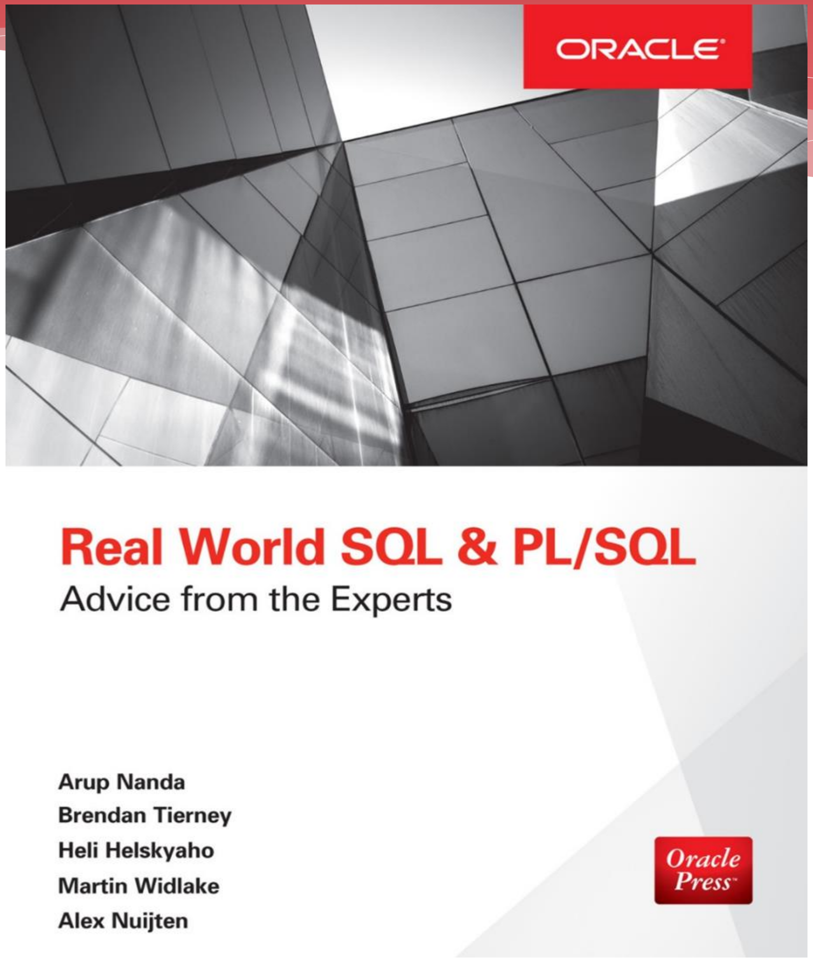



Miracle Finland Oy

# Predictive Queries in Oracle 12c

* Predictive Queries enable you to build and score data quickly using the in-database data mining algorithms
* Predictive Queries can be
  * built using Oracle Data Miner
  * written using SQL

# Chapter 12


ORACLE
Real World SQL & PL/SQL
Advice from the Experts
Arup Nanda
Brendan Tierney
Heli Helskyaho
Martin Widlake
Alex Nuijten
Oracle Press

# And so many more languages to learn…

* Python
* C/C++
* Java
* JavaScript
* Julia, Scala, Ruby, Octave, MATLAB, SAS

* https://medium.com/towards-data-science/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7

Miracle Finland Oy

# The future and now!

* AI and machine learning is here and it’s the future
* So many interesting areas to learn
* Pick your area and START LEARNING!

# Conclusions

* The time for Machine Learning is now because we technically able to use it and because of Big Data

# Conclusion

* Several V's related to Big Data…
  * Volume
  * Velocity
  * Variety
  * Veracity
  * Viability
  * Value
  * Variability
  * Visualization
  * …

MIRACLE
Miracle Finland Oy

# Conclusion

* ML can be used ”everywhere”:
    * Spam filters
    * Log filters (and alarms)
    * Data analytics
    * Image recognition
    * Speech recognition
    * Medical diagnosis
    * Robotics
    * …

MIRACLE
Miracle Finland Oy

# Conclusion

- Machine learning is all about approximation
- Unsupervised Learning vs supervised Learning
  - Unsupervised Learning
    - Clustering: hard or soft
  - Supervised Learning
    - Train, Predict
- Predictive Models:
    - classification, regression

# Conclusion

* Improving Models
  * Feature engineering
  * Hyperparameter tuning

* What to measure? How to interpret the measures?

* There is so much more to learn in ML...

Miracle Finland Oy

# THANK YOU!

QUESTIONS?

heli@miracleoy.fi

Twitter: @HeliFromFinland

Blog: Helifromfinland.com

Miracle Finland Oy