Using Machine Learning to unlock the Business Value in Big Data

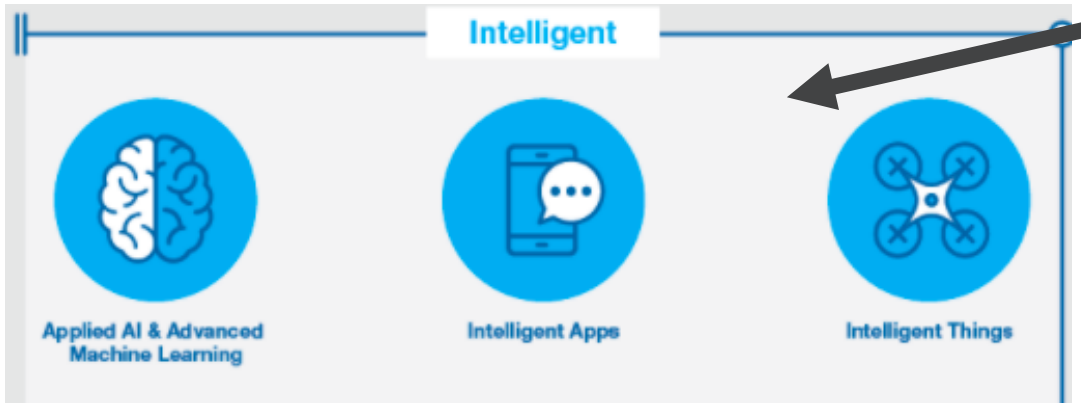
Data Science for Big Data and Cloud

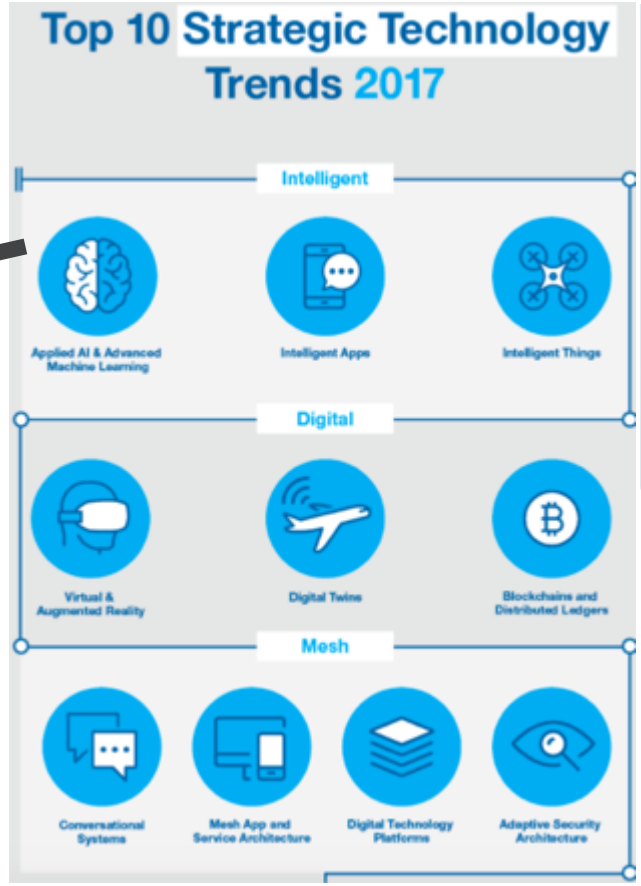
Marcos Arancibia
Product Manager, Oracle Data Science

Gartner's Top 10 Strategic Technology Trends for 2017

Artificial intelligence, machine learning, and smart things promise an intelligent future.

October 18, 2016
Contributor: Kasey Panetta





#1 - Applied AI & Advanced Machine Learning

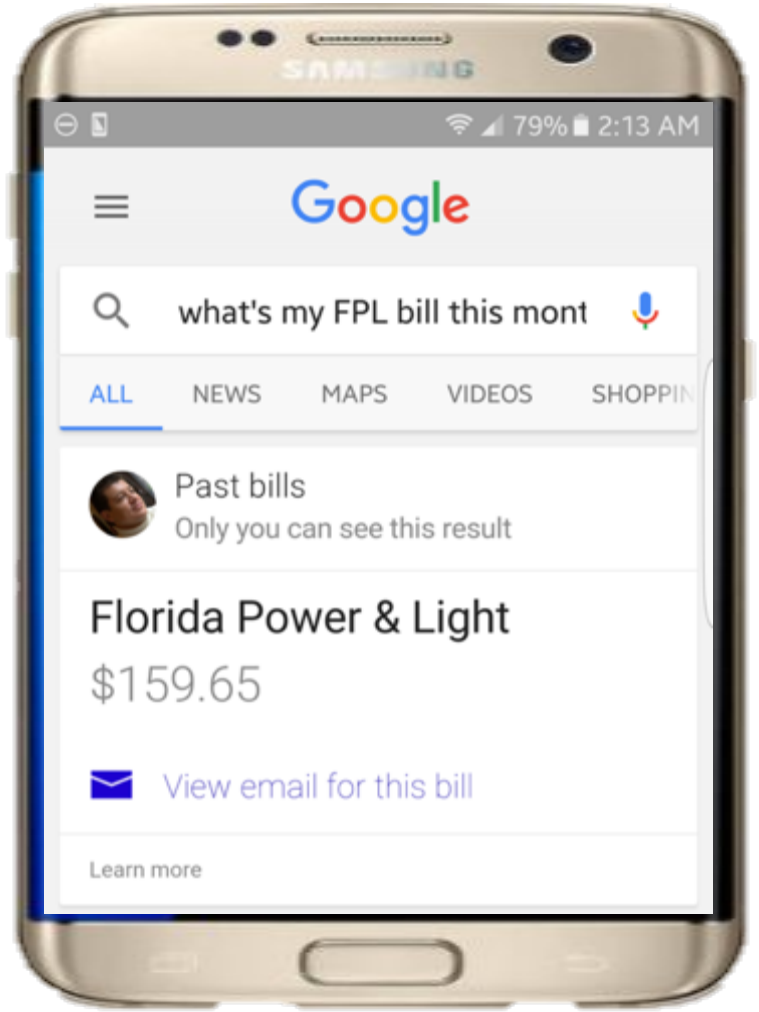
#2 – Intelligent Apps
#3 – Intelligent Things

Source: http://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017/

What is Machine Learning?

And how can it benefit my business?

Machine Learning – Useful or Scary?



What's happening in my Phone when I ask it what's my Electricity bill?

1. Speech recognition using an API
2. Massive contextual search to understand what it means remotely (NLP)
3. Search through my personal e-mails (thanks to Gmail)
4. More context detection to find the exact value of the Bill
5. Return to the phone, speaking out loud the value using text-to-speech
6. Provide the link to the original e-mail

Types of uses of ML by Data Warehouse Oracle customers

What can Machine Learning do for our customers?

- You know your customer e-mail preferences and some purchase habits, but can you really tell this group of customers apart?
- Do you know how likely each one is to accept the next offer before blasting their inboxes?
- How can you understand their similarities and differences?
- How to differentiate good from bad transactions from these customers?

Types of uses of ML by Data Warehouse Oracle customers

Better business through Intelligent Customer Targeting

- Machine Learning techniques bring advanced statistical and mathematical analysis to the Enterprise level to help understand customer behavior.
- Classification models help us identify the best targets for each unique campaign type, and anticipate customer reactions.
- Segmentation models helps us understand the different customer needs, including their Purchasing behavior (not just the traditional Age, gender and geographical locations).
- Fraud Detection models will stop Fraud before it happens, or flag a transaction for further review

Types of uses of ML by Data Warehouse Oracle customers

How to best classify your customers?

?

Features:

Basic Query	Age/Gender	Known	Known	Known	Unknown	Unknown
	Marketing Preferences	Mail and e-mail	e-mail	e-mail and Facebook	e-mail and Google+	Mail, e-mail and Twitter

Types of uses of ML by Data Warehouse Oracle customers

How to best classify your customers?

?

	Features:					
Basic Query	Age/Gender	Known	Known	Known	Unknown	Unknown
	Marketing Preferences	Mail and e-mail	e-mail	e-mail and Facebook	e-mail and Google+	Mail, e-mail and Twitter
Traditional; Analytics	RFM (Recency, Frequency and Monetary Value): Purchases in the last 3/6/12 mo.	1 item/$35 in the last 3 mo	2 items/$150 in the last 6 mo	3 items/$75 in the last 3 mo	3 items/$225 in the last 12 mo	9 items/$250 in the last 6 mo

Types of uses of ML by Data Warehouse Oracle customers

How to best classify your customers?

?

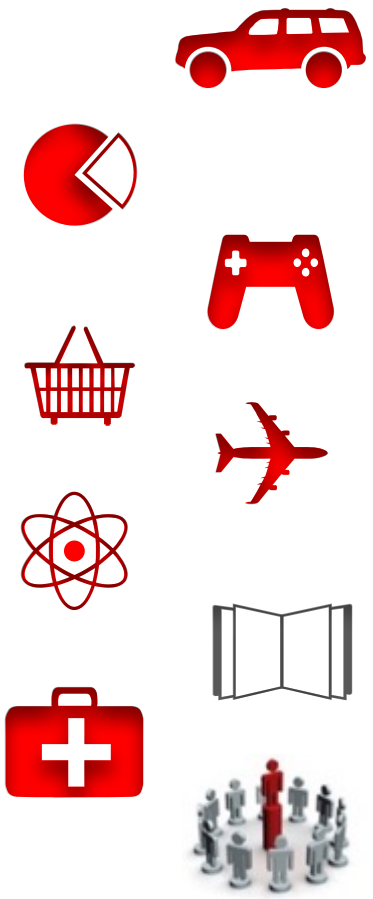
Features:					
Basic Query	Age/Gender	Known	Known	Known	Unknown	Unknown
	Marketing Preferences	Mail and e-mail	e-mail	e-mail and Facebook	e-mail and Google+	Mail, e-mail and Twitter
Traditional; Analytics	RFM (Recency, Frequency and Monetary Value): Purchases in the last 3/6/12 mo.	1 item/$35 in the last 3 mo	2 items/$150 in the last 6 mo	3 items/$75 in the last 3 mo	3 items/$225 in the last 12 mo	9 items/$250 in the last 6 mo
Machine Learning	**Behavioral Customer Segment**	**"Retired Cosmopolitan"**	**"Affluential Executive"**	**"New Home Mom"**	**"Young Successful startup"**	**"Executive product collector"**
	Probability to buy New Product X	**31%**	**45%**	**55%**	**21%**	**72%**

ORACLE

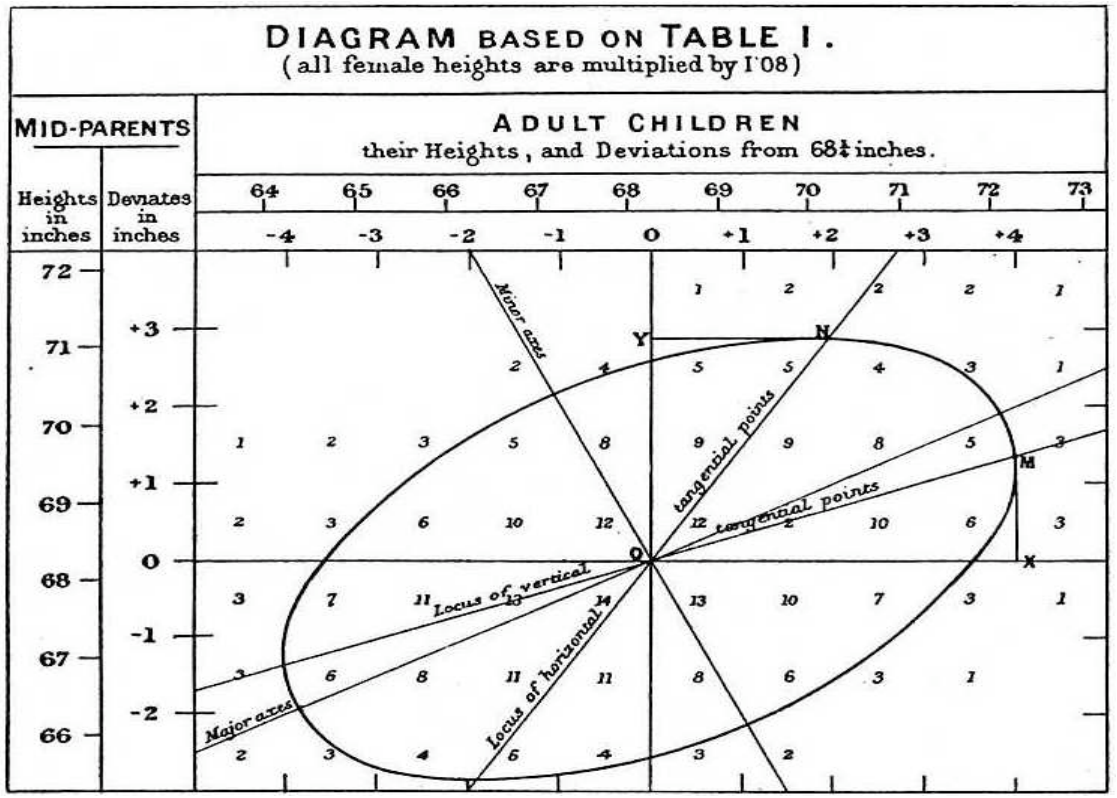
Types of uses of ML by Data Warehouse Oracle customers

Typical use case scenarios

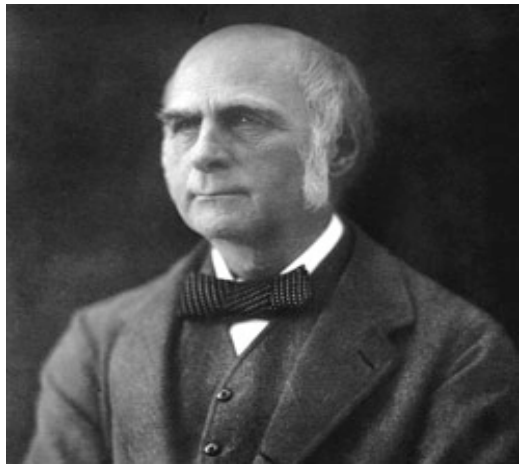
- Targeting the right customer with the right offer
- How is a customer likely to respond to an offer
- Finding the most profitable growth opportunities and customers
- Finding and preventing customer churn
- Maximizing cross-business impact
- Security and suspicious activity detection
- Understanding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



Linear Regression history

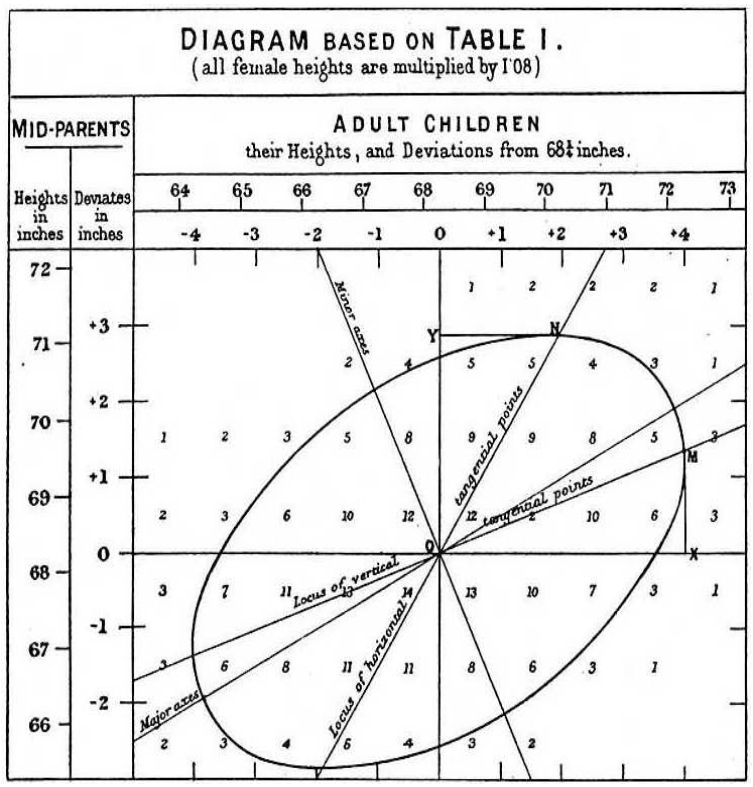


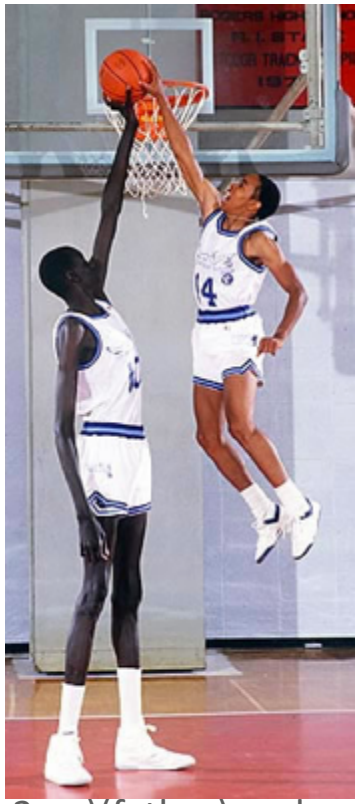
Sources: http://en.wikipedia.org/wiki/Francis_Galton



Francis Galton (half-cousin of Charles Darwin) was the first to describe and explain the common phenomenon of **Regression** toward the mean, which he first observed in his experiments on the size of the seeds of successive generations of sweet peas in 1875.

Linear Regression history



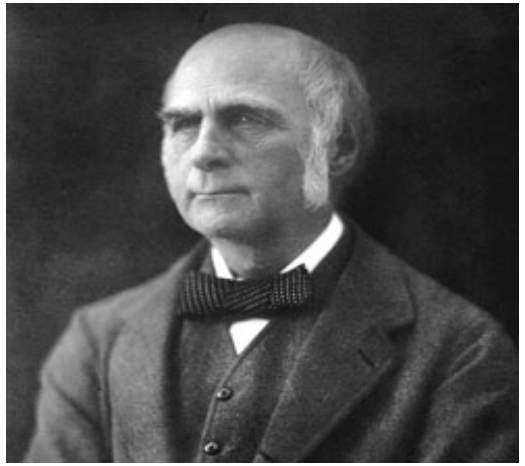


Manute Bol was 7'7"(2m 31cm) to a 6'8" (2m 3cm)(father) and a 6'10" (2m 8cm) (mother)…

Regressed to

but his grandfather was 7'4" (2m 24cm)

Source: http://en.wikipedia.org/wiki/Manute_Bol

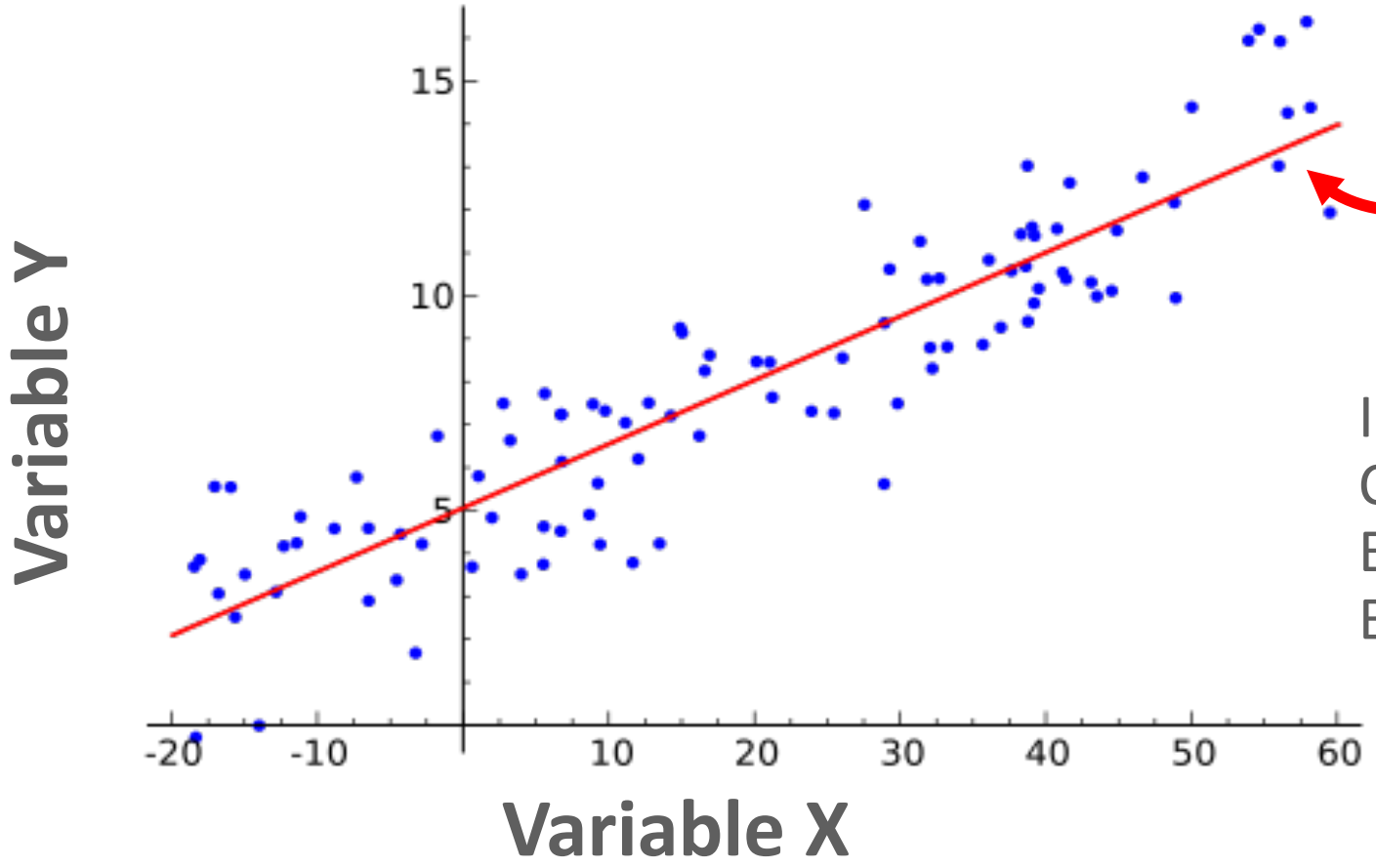


Francis Galton (half-cousin of Charles Darwin) was the first to describe and explain the common phenomenon of **Regression** toward the mean, which he first observed in his experiments on the size of the seeds of successive generations of sweet peas in 1875.

Source: http://en.wikipedia.org/wiki/Francis_Galton

Linear Regression

Many names from different disciplines



Dependent Variable,
Target Attribute

Explanatory Variable,
Input Attribute

$$Y = a + bX$$

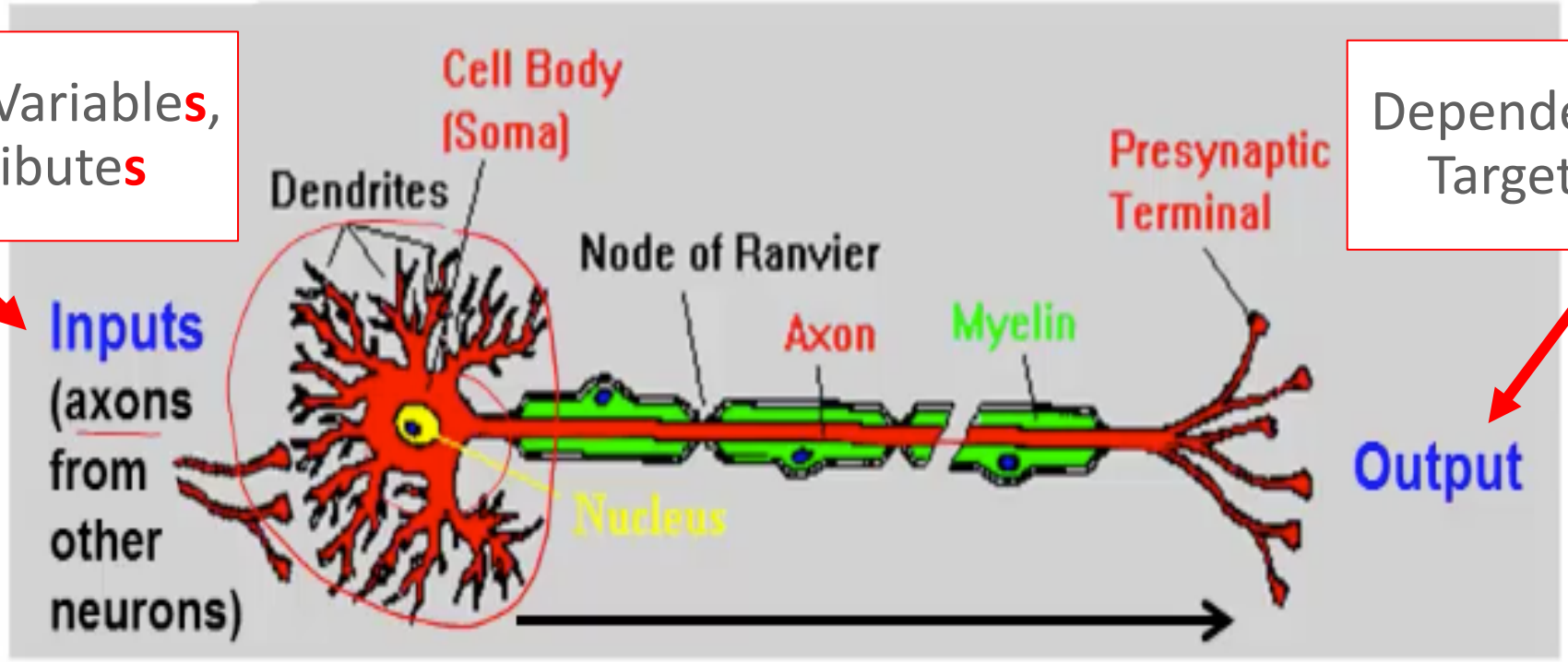
Intercept,
Constant,
Error,
Bias

Coefficient,
Weight,
also intuitively:
”Correlation”
between
attributes

https://en.wikipedia.org/wiki/Linear_regression

The idealized Neuron

Explanatory Variables, Input Attributes



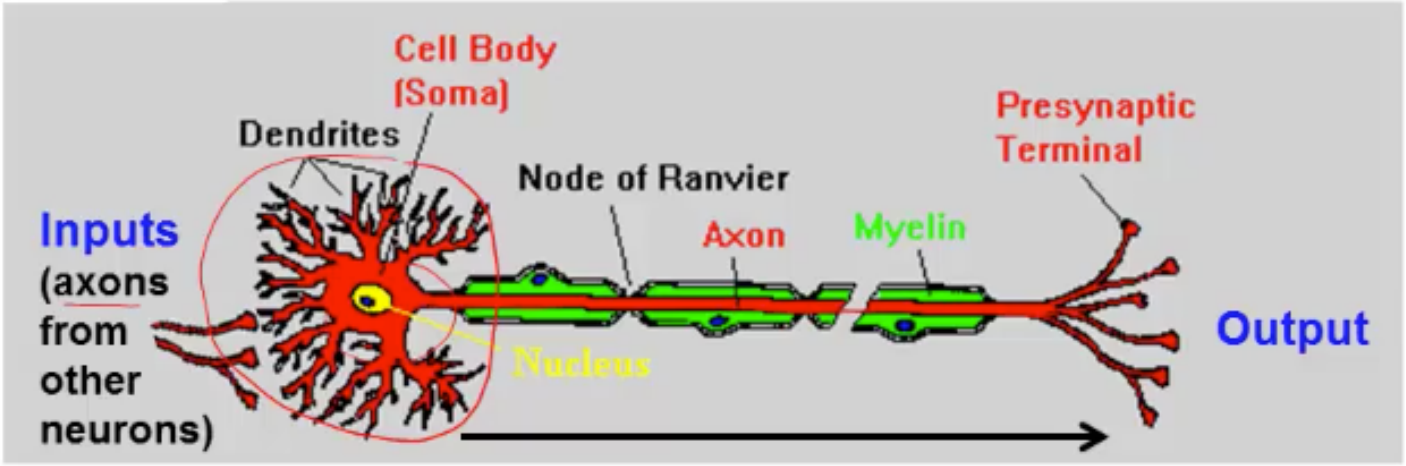
Dependent Variable, Target Attribute

https://www.coursera.org/learn/computational-neuroscience/lecture/iynBe/1-4-the-electrical-personality-of-neurons

The idealized Neuron in practice

INPUT

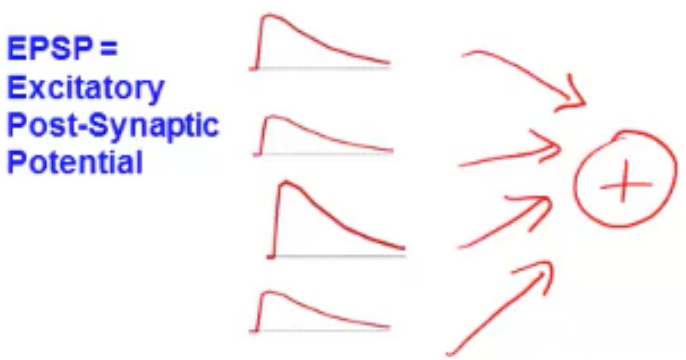
OUTPUT

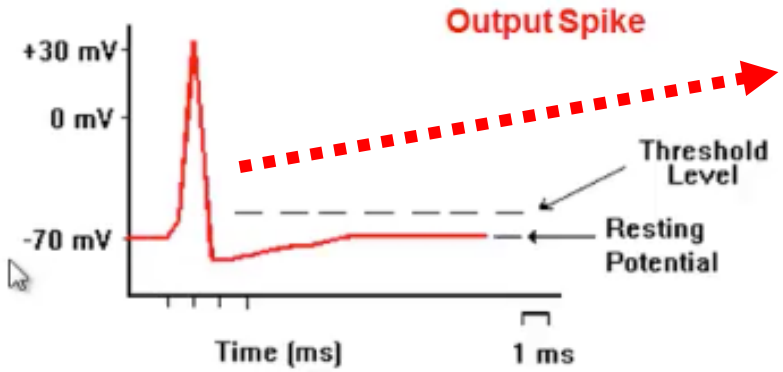


Years in Job

Income

etc…

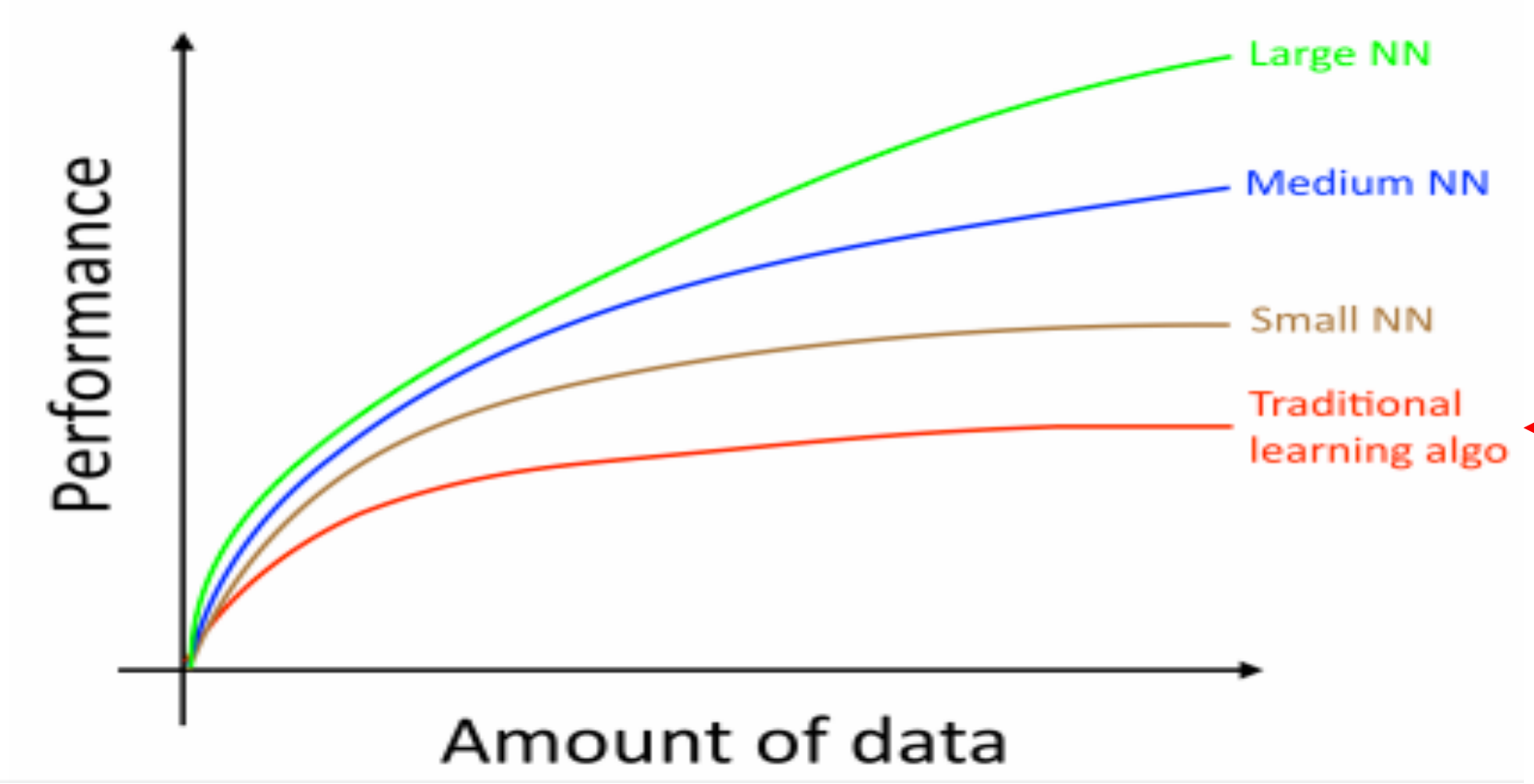




PURCHASE INSURANCE

https://www.coursera.org/learn/computational-neuroscience/lecture/iynBe/1-4-the-electrical-personality-of-neurons

Deep Learning and Deep Neural Networks
Motivation for the future

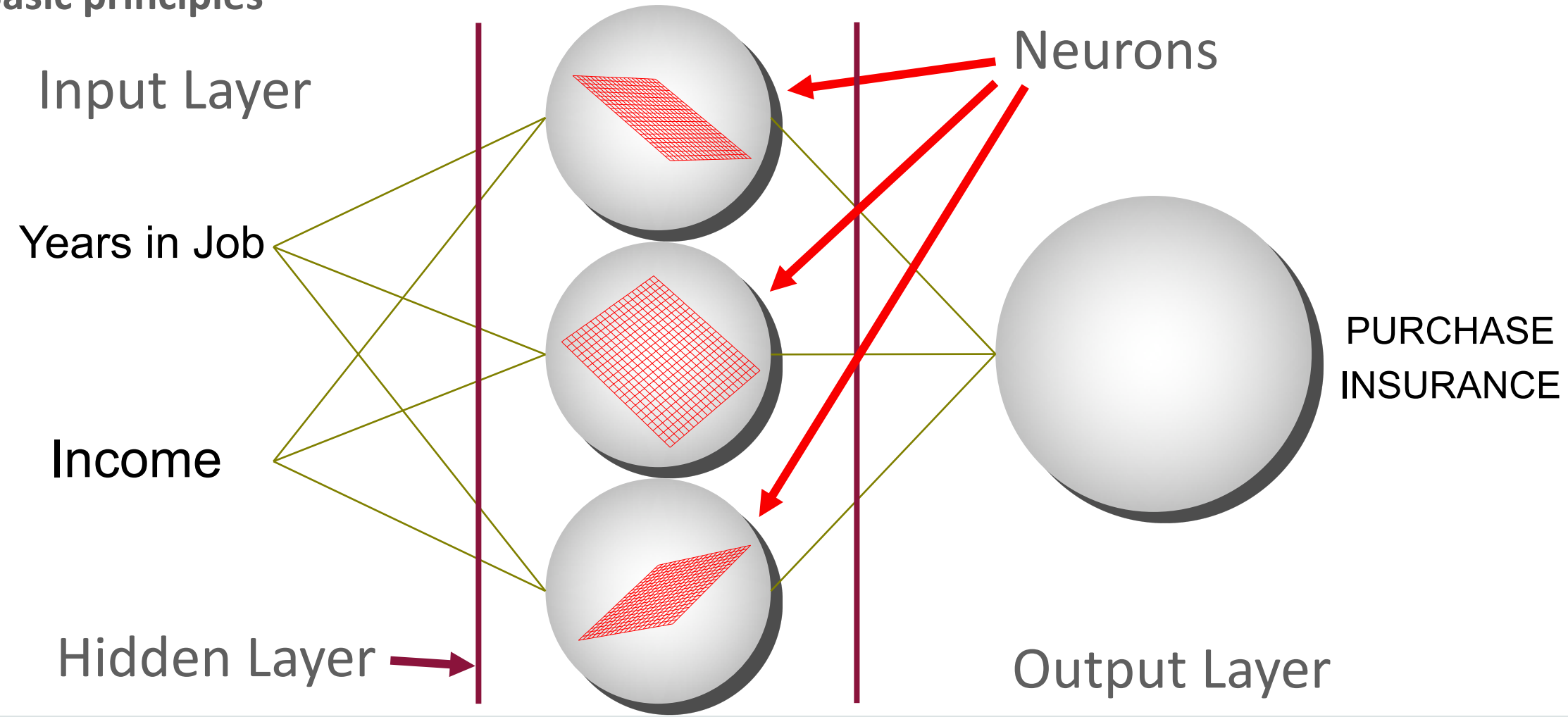


Regression,
Decision Trees,
other traditional algorithms

Source: https://gallery.mailchimp.com/dc3a7ef4d750c0abfc19202a3/files/Machine_Learning_Yearning_V0.5_01.pdf

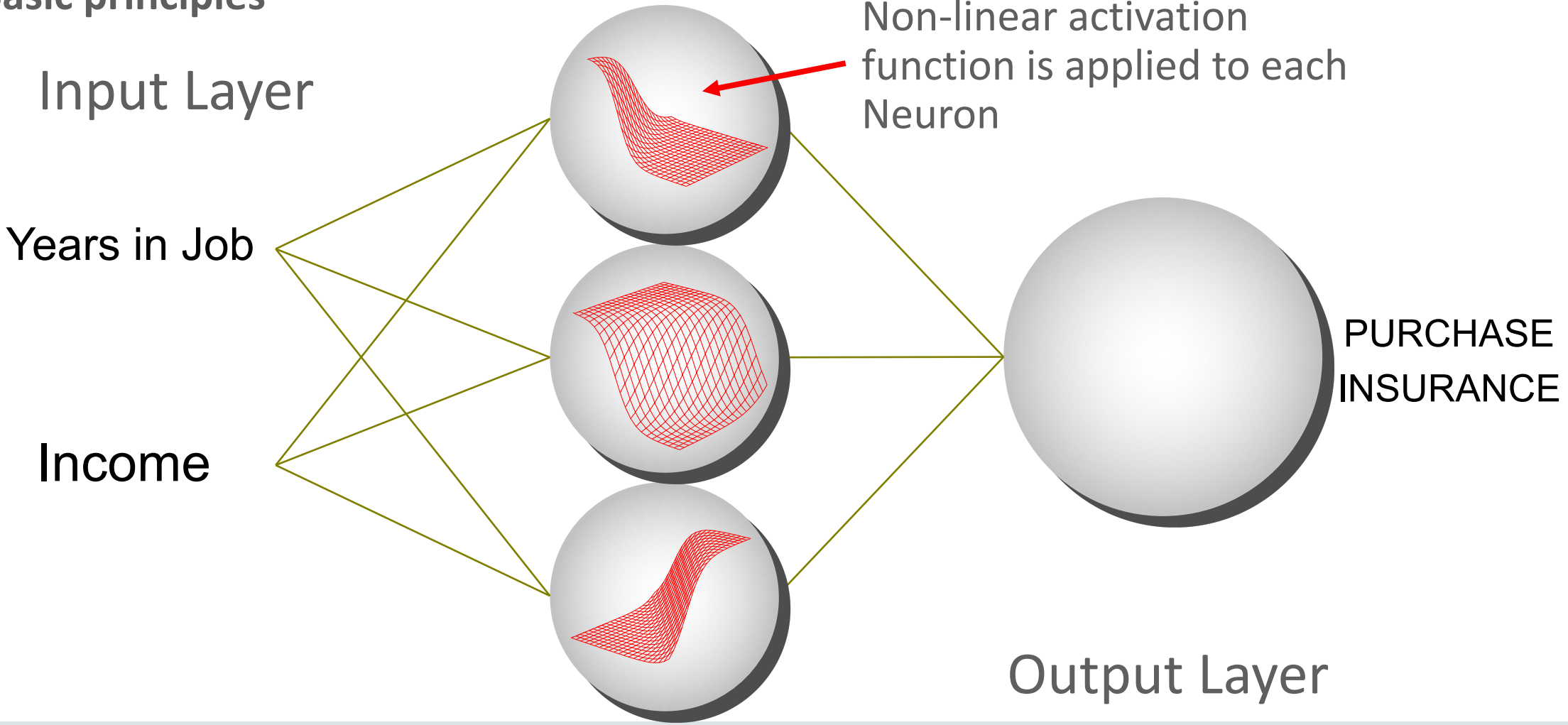
Machine Learning Basics - Neural Networks – MLP

Basic principles



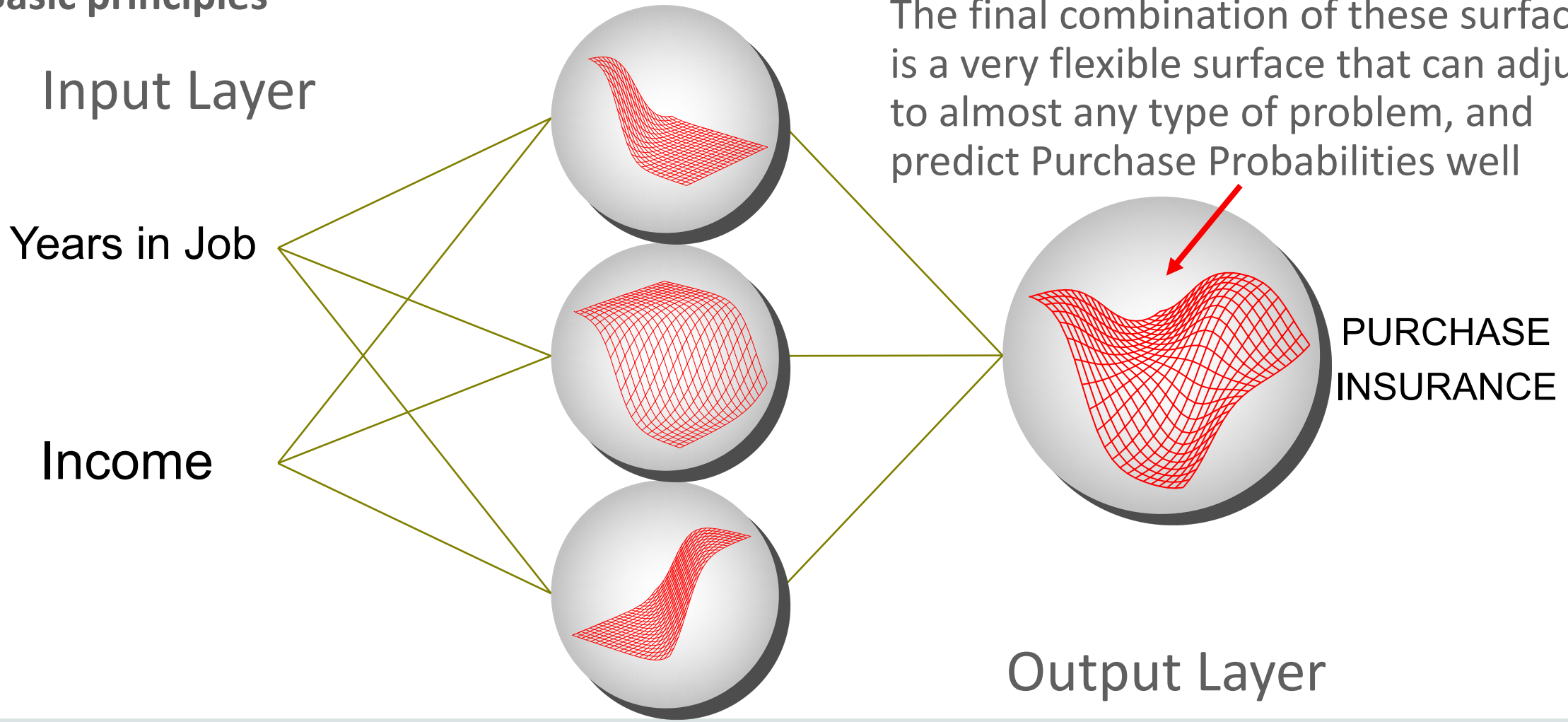
Machine Learning Basics - Neural Networks – MLP

Basic principles



Machine Learning Basics - Neural Networks – MLP

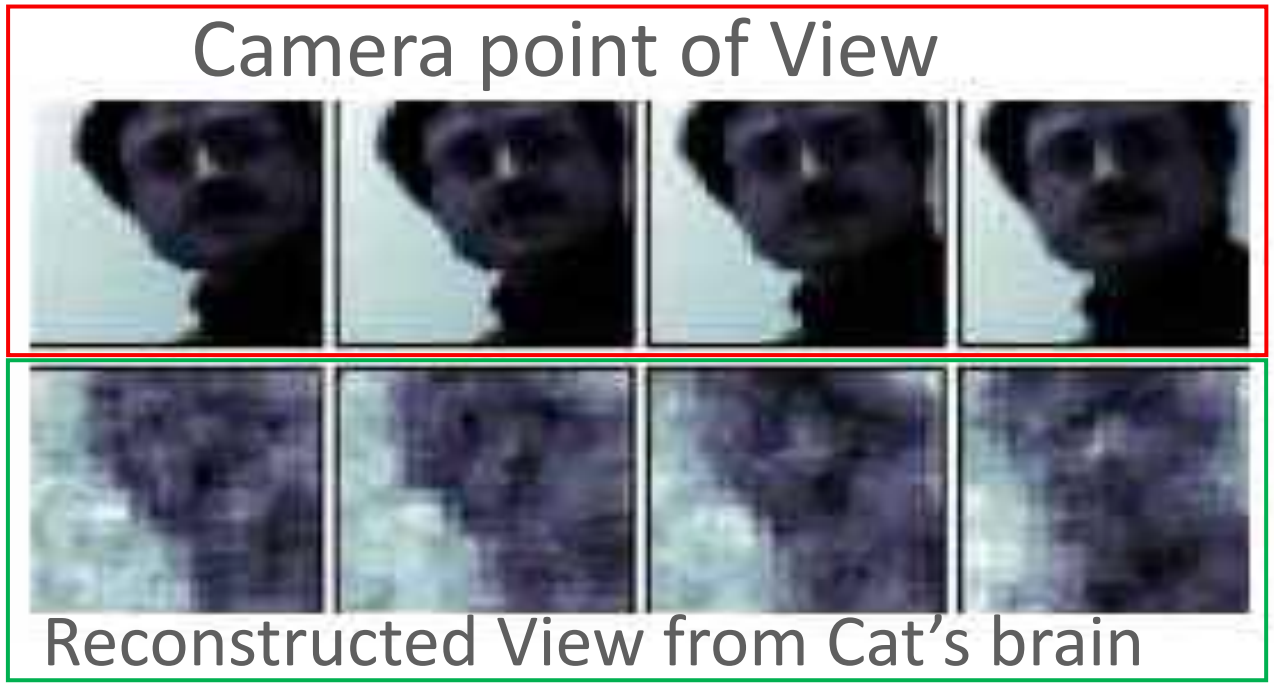
Basic principles



Experiments with Neural Networks in Neuroscience

INCORPORATED ANYONE ?

In 1999, Scientists at the University of California, Berkeley, led by Dr. Yang Dan, assistant professor of neurobiology, have recorded signals from deep in the brain of a cat to capture movies of how it views the world around it



"…Cat's brain signals from a total of 177 cells in the lateral geniculate nucleus (a part of the brain's thalamus) that processes visual signals from the eye…"

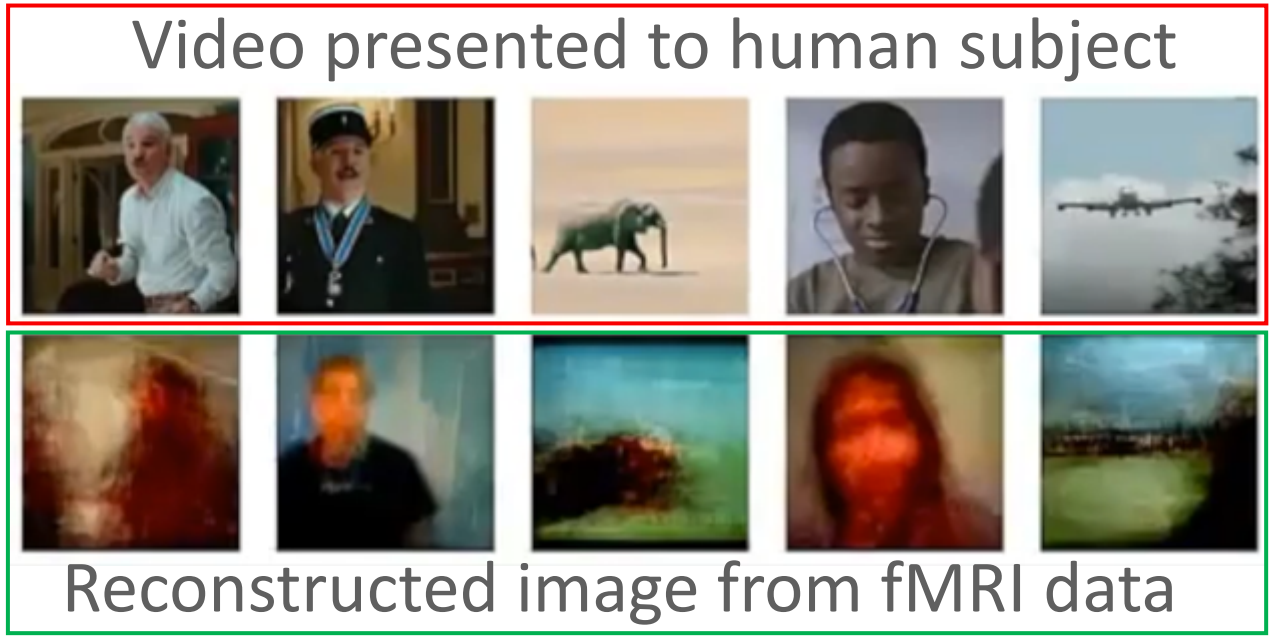
http://www.berkeley.edu/news/media/releases/99legacy/10-15-1999.html

http://www.coursera.org/learn/computational-neuroscience/

Experiments with Neural Networks in Neuroscience

INCORPORATED ANYONE ?

In 2011, Nishimoto et al. published an article on *Current Biology* on the first reconstructions of natural movies from human brain activity. This is a critical step toward the creation of brain reading devices that can reconstruct dynamic perceptual experiences.

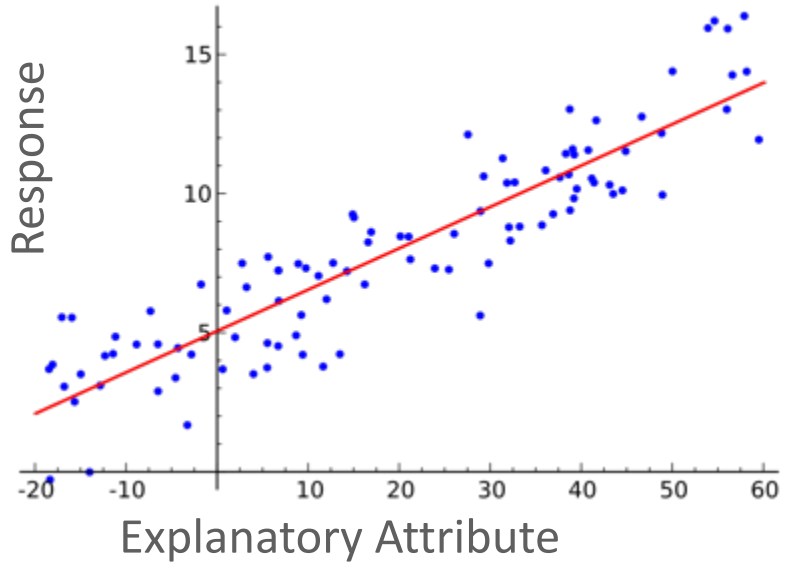


"…This modeling framework might also permit reconstruction of dynamic mental content such as continuous natural visual imagery…and could potentially be used to decode involuntary subjective mental states (e.g., dreaming or hallucination)…"

http://www.cell.com/current-biology/fulltext/S0960-9822(11)00937-7#
http://www.coursera.org/learn/computational-neuroscience/

Machine Learning algorithms

Supervised Learning Regression Problem types and applicability

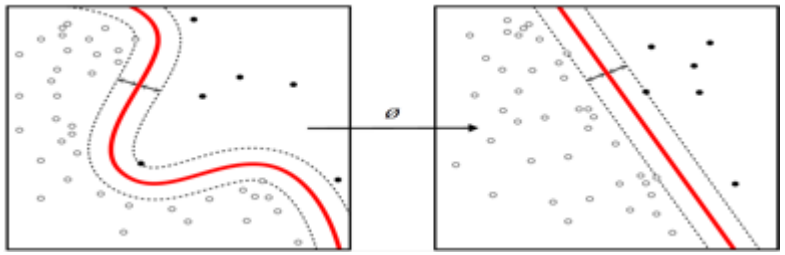
Problem Type	Algorithms	Applicability
Regression 	• Multiple Regression (GLM) • Support Vector Machine • Stepwise Linear Model • Deep Neural Networks • Random Forest • LASSO • Ridge Regression	• Predict Usage • Estimate Credit Limits • Estimate optimal Pricing

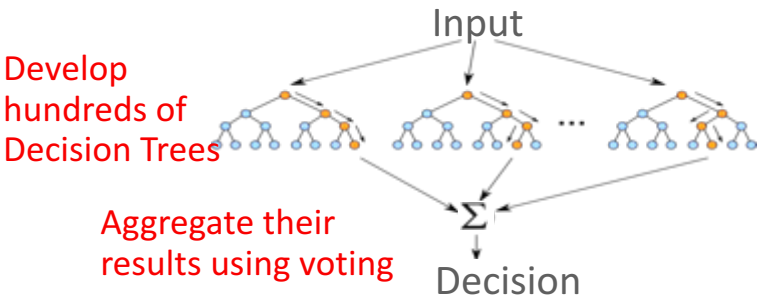
https://commons.wikimedia.org/wiki/File:Linear_regression.svg

Machine Learning algorithms

Supervised Learning Classification Problem types and applicability

Problem Type	Algorithms	Applicability
Classification	• Logistic Regression (GLM) • Decision Trees • Support Vector Machines • Deep Neural Networks • Random Forest	• Probability to Churn • Probability to Purchase • Likelihood of Fraud





https://en.wikipedia.org/wiki/Supervised_learning#/media/File:Kernel_Machine.svg

Customer Experiences with cost and time savings

- StubHub reduced Fraud on the credit card ticket sales by 95% with the use of Oracle R Enterprise Models stored directly in their e-Commerce Database
- Models are moved from Dev to Production within the same day, saving a lot of time and avoiding great losses
- Enabled data scientists to use R and avoid translation for production deployment – same day deployment

Don’t need to move the data around

- Wargaming.net, an online game developer and leading publisher in the MMO game market, is able to use the Telemetry of more than 110M gamers and up to 300B events to run Machine Learning Models
- Reduced execution time of complex models from 6 hours to just 3 minutes with Oracle Advanced Analytics
- Increased revenues in certain segments by 62% with targeted campaigns

Massive scalability

- With 25% market share, ZABA is the leading Croatian financial institution.
- Migrated from SAS to OAA and increased performance. SAS took days to complete, vs. minutes on OAA.
- Saved 1,000 person-days/year in IT and increased Cash Loans by 15% in 18 months due to improved hit ratios
- Historical customer behavior analysis shortened from several months to 2 weeks

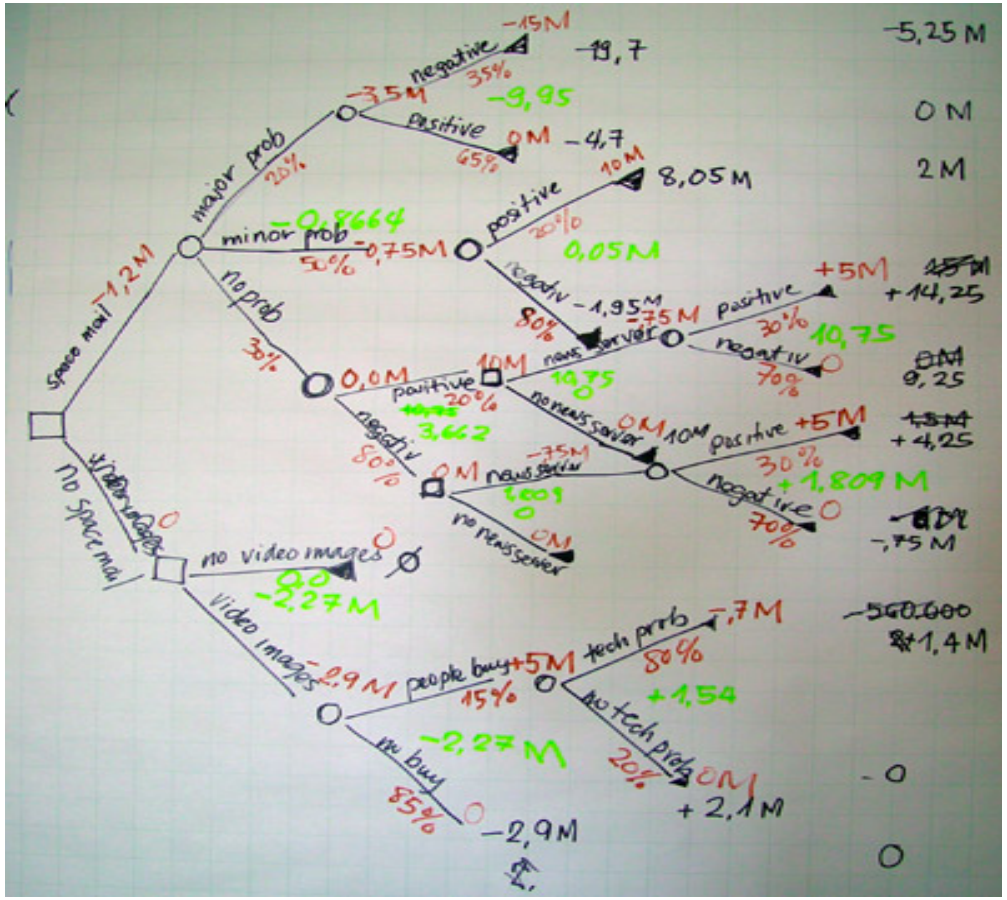
Time to market

Machine Learning Basics - Decision Trees

Usage in Marketing, CRM, Risk Models

Typical algorithms:

- CART® – Classification and Regression Trees
  - First introduced by by Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen in 1984
- CHAID – Chi-Squared Automatic Interaction Detection
- ID3, C4.5/C5, MARS
- Random Forest
- Gradient Boosted Trees
- many more…

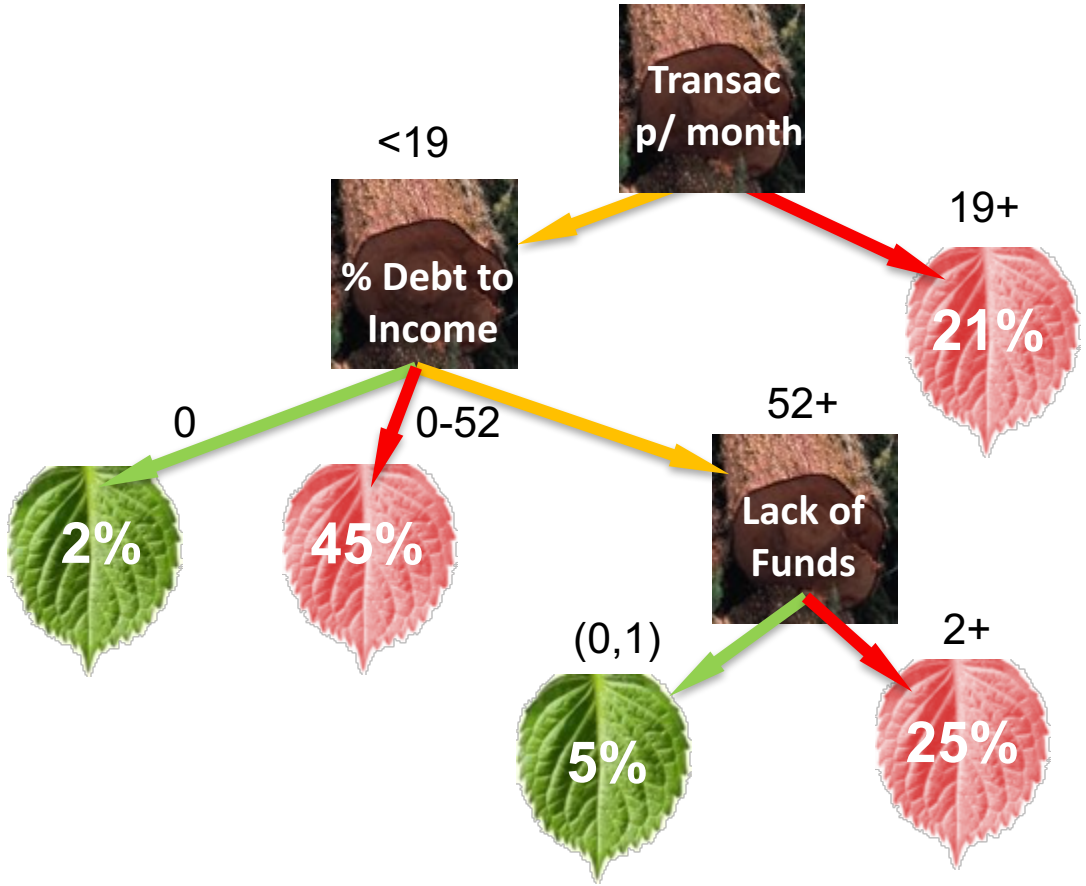


Source: https://en.wikipedia.org/wiki/Decision_tree

Machine Learning Basics - Decision Trees - Advantages

Usage in Marketing and CRM Models

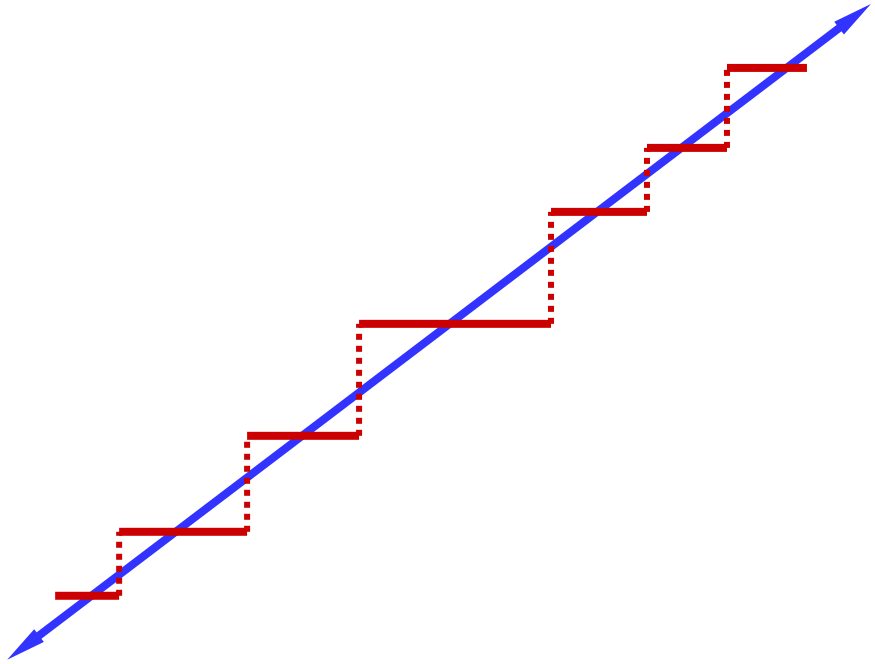
- Interpretability
- Tree structured presentation
- Mixed Measurement Scales
- Nominal, ordinal, interval
- Regression trees
- Robustness
- Native Support for Missing Values



Machine Learning Basics - Decision Trees - Disadvantages

Usage in Marketing and CRM Models

- Roughness – Step ladder
- Linear, Main Effects by default
- Instability – what happens on your birthday if the model uses Age?

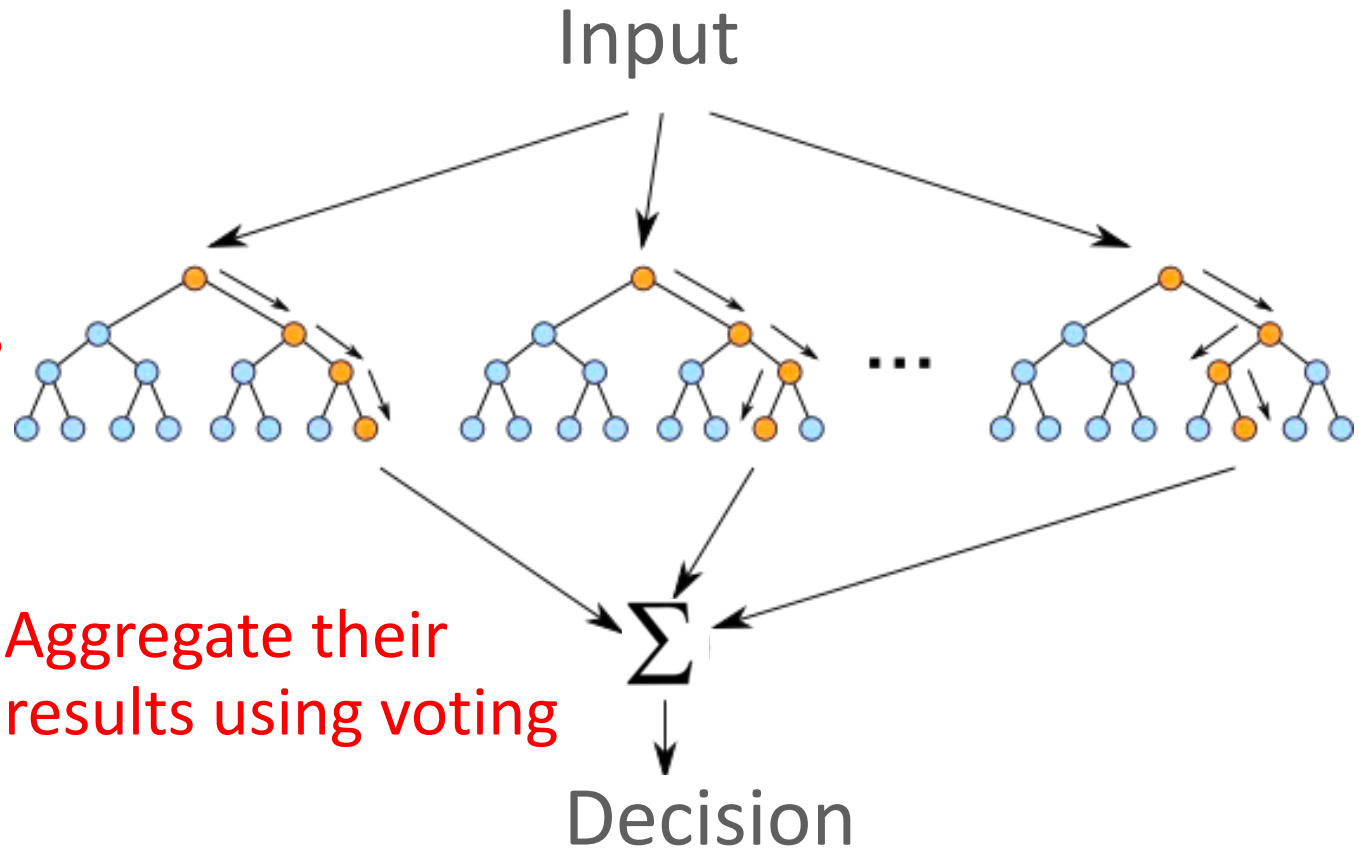


Machine Learning Basics - Decision Trees - Ensembles

Usage in Marketing and CRM Models

- Random Forests
- Gradient Boosted Trees

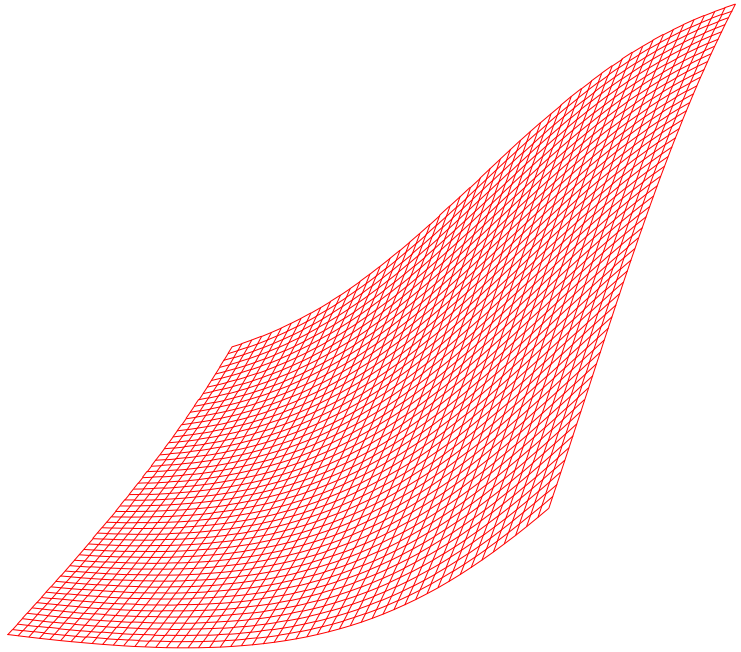
Develop hundreds of Decision Trees



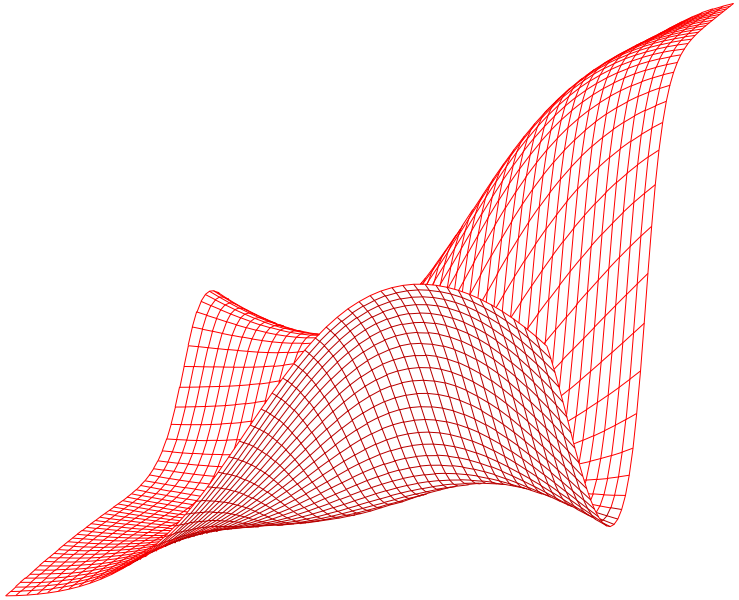
Aggregate their results using voting

Machine Learning Basics – Methods compared

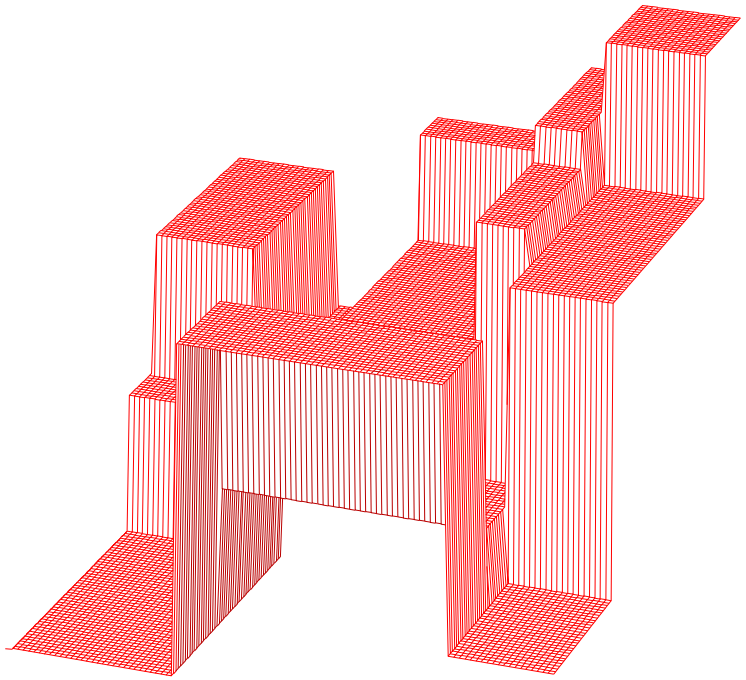
Comparing a few of the different analytical methods



Generalized
Linear Models



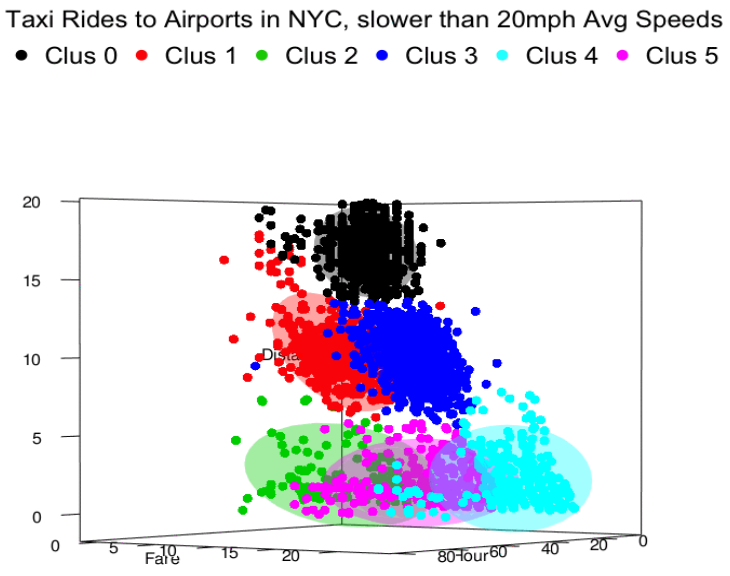
Neural
Networks,
SVM



Decision
Trees

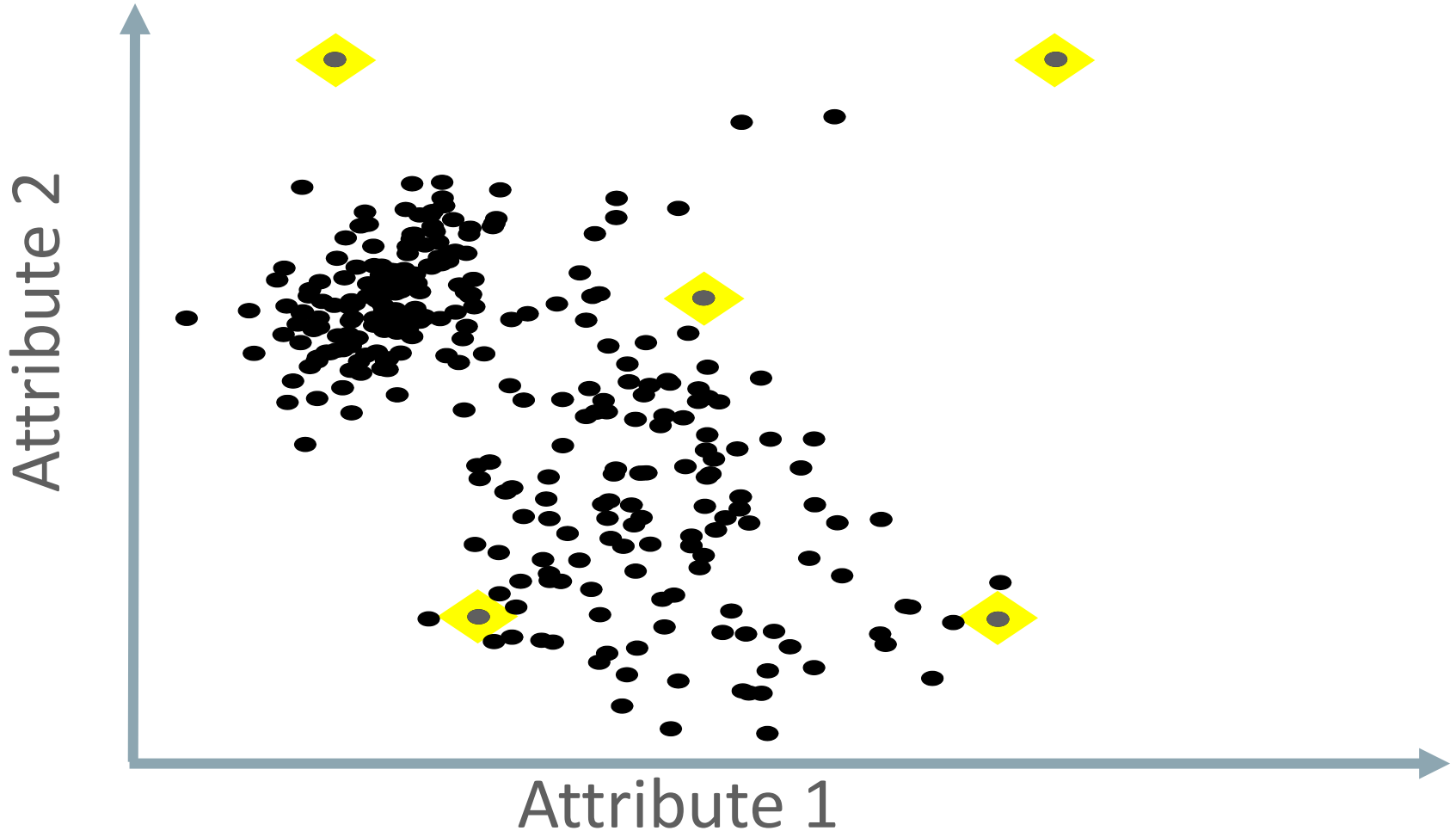
Machine Learning algorithms

Unsupervised Learning Problem types and applicability

Problem Type	Algorithms	Applicability
Clustering 	• Hierarchical K-Means • Hierarchical O-Cluster • Expectation-Maximization Clustering • Gaussian Mixture Model	• Group Customers in Similar Behaviors • Product Grouping • Text Mining

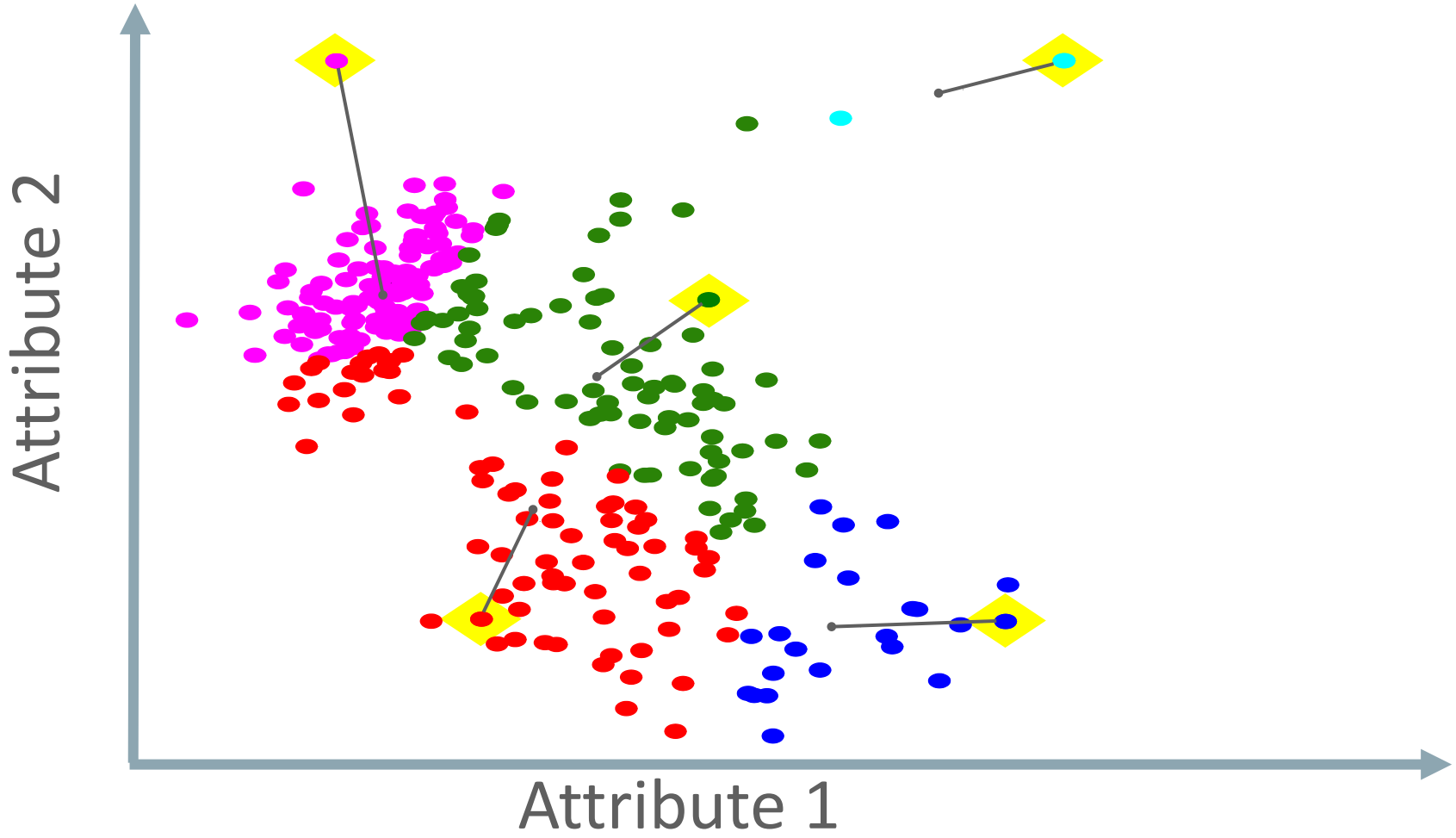
Machine Learning Basics - Iterative Clustering Models

k-Means Applied to Segmentation – Phase 1 – Random Pick of k points



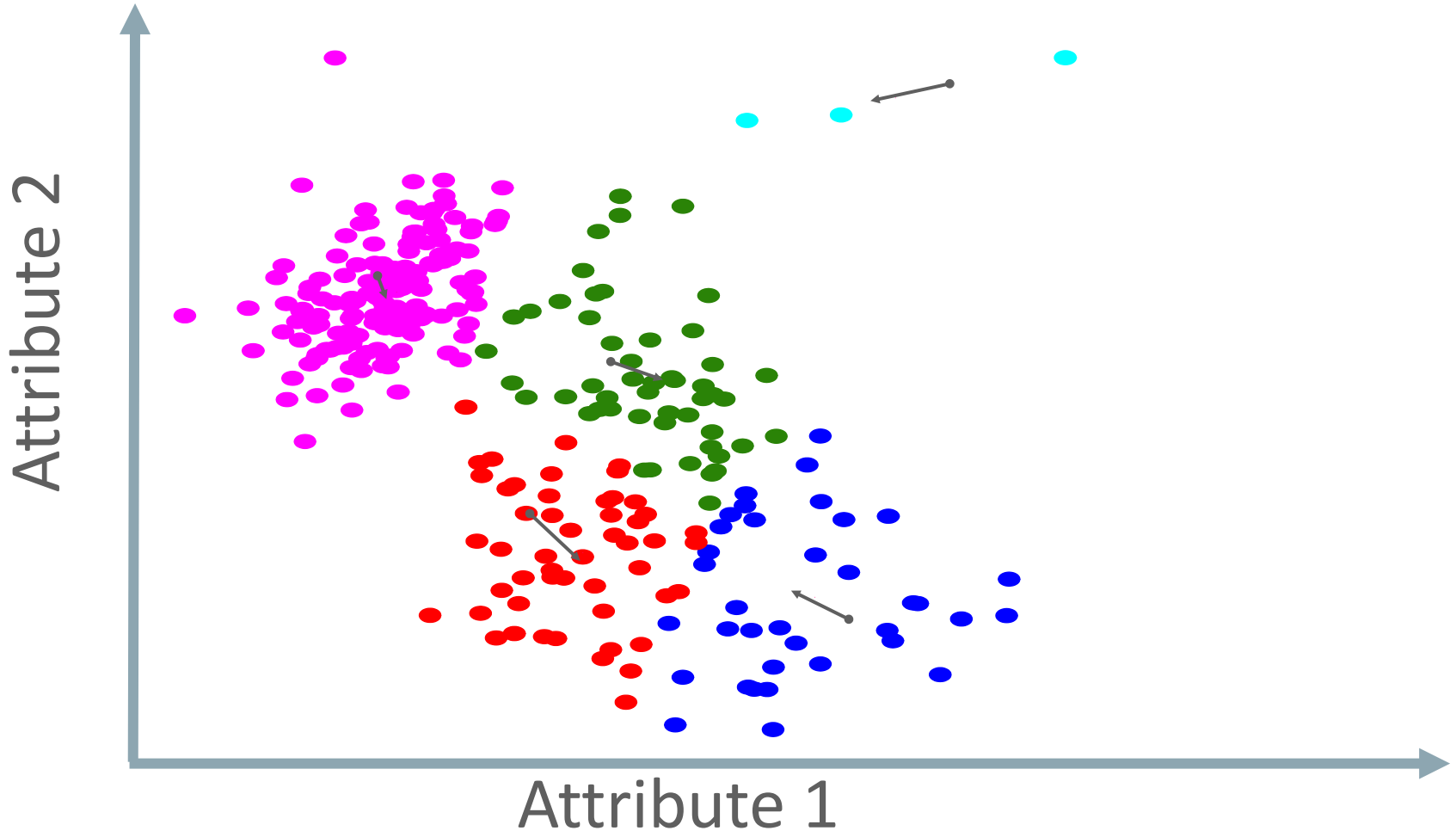
Machine Learning Basics - Iterative Clustering Models

k-Means Applied to Segmentation – Phase 2 – Cluster Assignment



Machine Learning Basics - Iterative Clustering Models

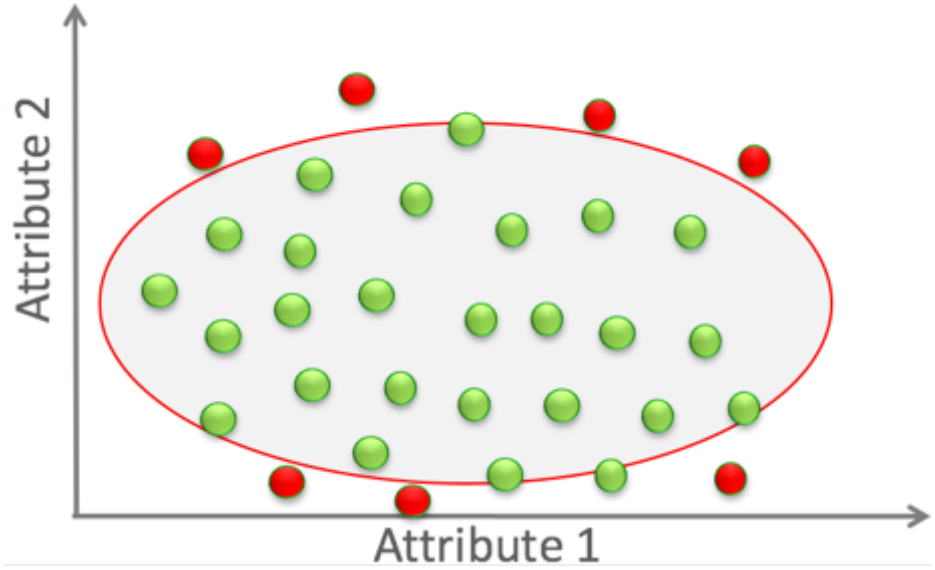
k-Means Applied to Segmentation – Phase 3 – Reassignment after re-center



Machine Learning algorithms

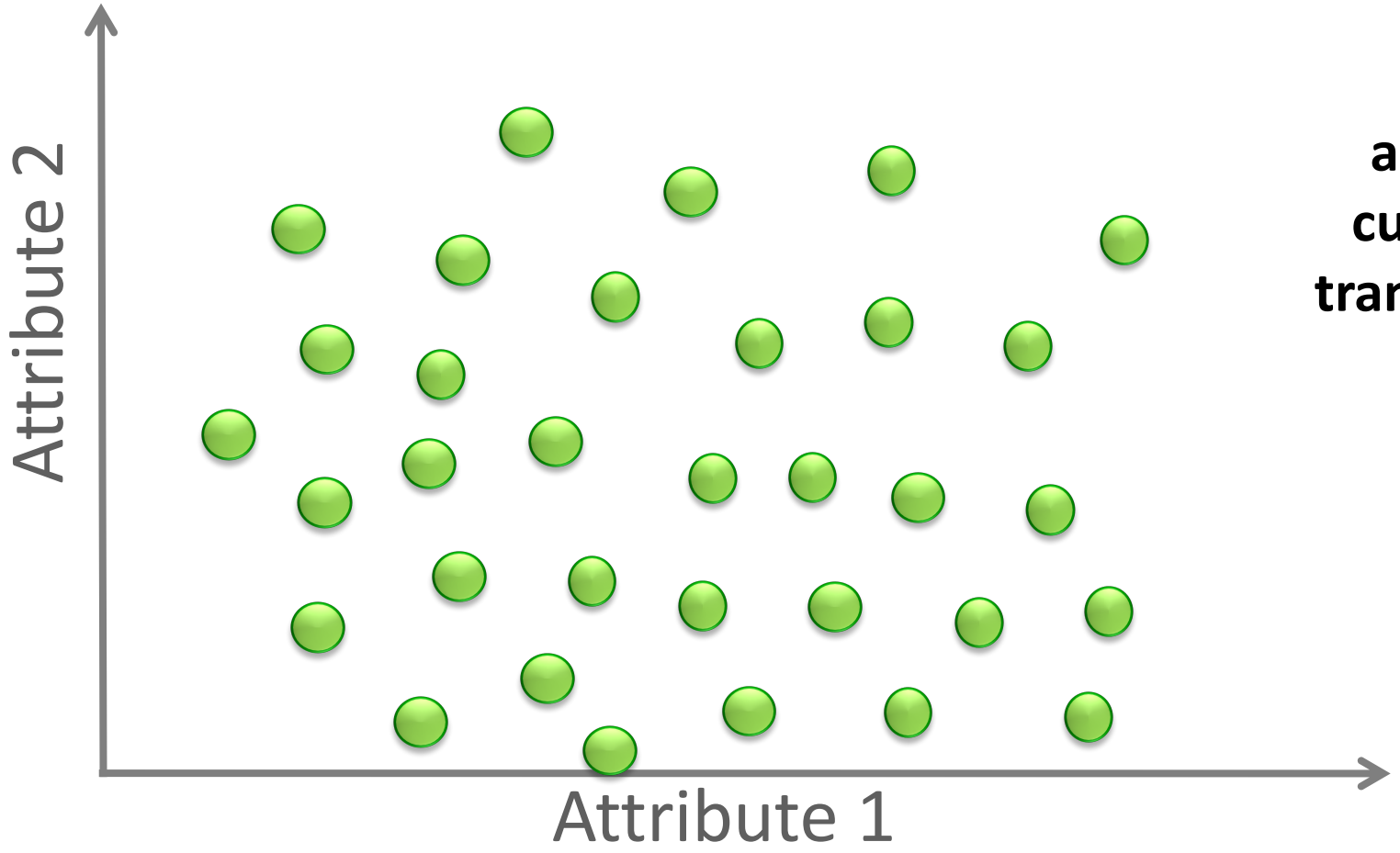
Unsupervised Learning Problem types and applicability

Problem Type	Algorithms	Applicability
Anomaly Detection	• One-Class SVM	• Identify anomalies or potential Fraud cases • Identify Intrusion on Networks based on uncommon behavior



Machine Learning Basics - Anomaly Detection Models

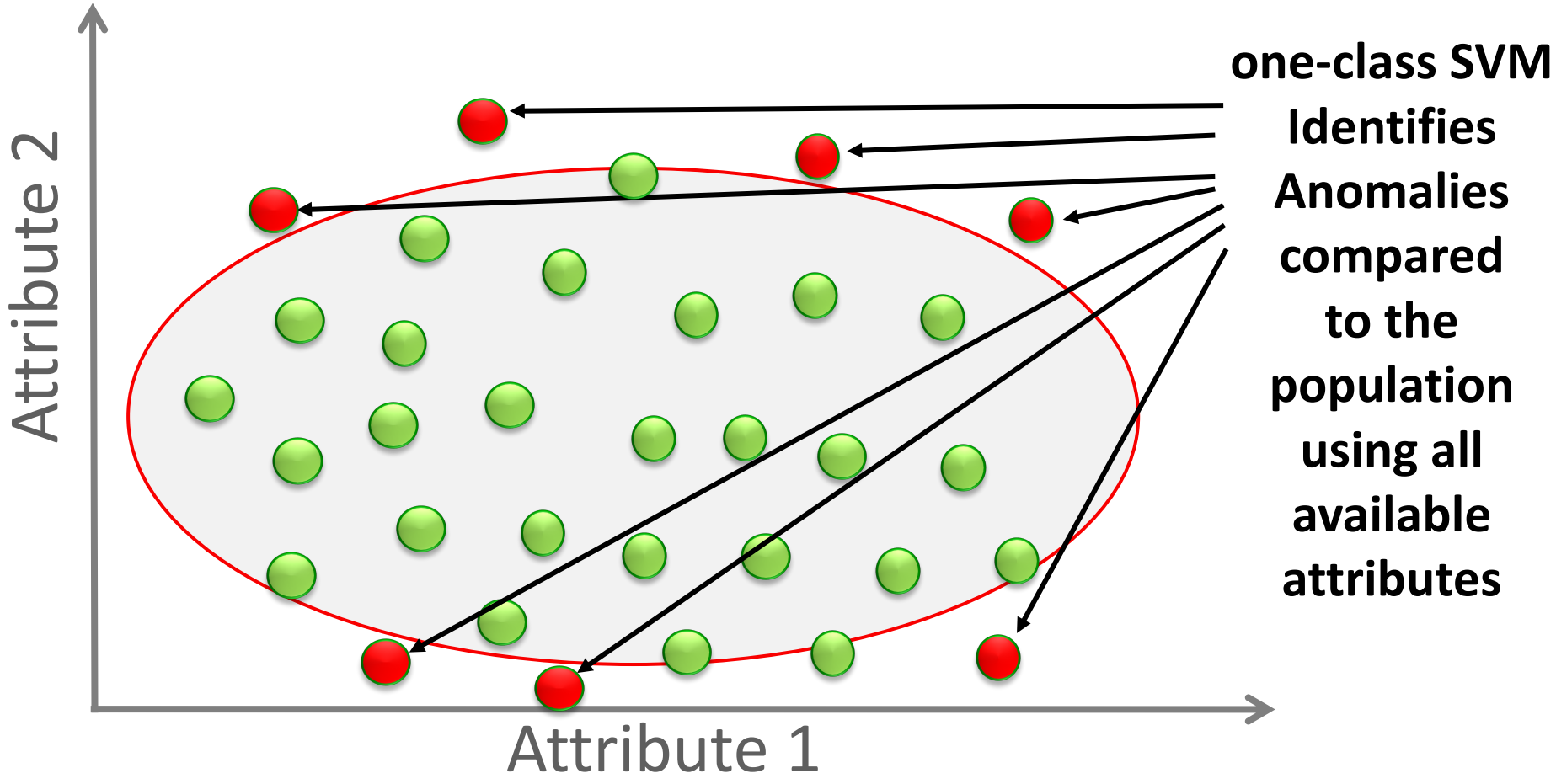
Unsupervised version – when you don't have examples of bad customers



Initially assumes all customers or transactions are normal

Machine Learning Basics - Anomaly Detection Models

Unsupervised version – when you don't have examples of bad customers



Machine Learning by Oracle

Vision and Architecture

Vision

- Big Data + Data Science Platform for the Era of Big Data and Cloud
  - Make Big Data + Machine Learning Model Building *Simple*
  - Make Big Data + Machine Learning Model Deployment *Simple*
  - *Key Differentiators:*
    - Fully integrated into Oracle Database and Hadoop platforms
    - Scalable and Distributed algorithms run where the data is
    - Easy to use using familiar interfaces like R.
    - Support for open-source R packages running in the Database Server or in the Cluster
    - Integrated with Oracle solutions and Cloud Services like Graph, OBIEE, BDD, IoT, OSA, RTD
    - Compatible with 3rd party GUIs like Rstudio, Jupyter and Zeppelin
    - Low TCO, included in Oracle Cloud Services like BDCS, DBCS and DBC-EXADATA Service

Oracle is a founding member of the R consortium

- **R Consortium Central mission** – work with and provide support to the **R Foundation** and R Community including key organizations developing, maintaining, distributing and using R software through the identification, development and implementation of infrastructure projects
- Enable the R user community to grow without disrupting R language development or the work of the R Foundation
- Organized under an open source governance and foundation model
  - Consists of Board of Directors, Infrastructure Steering Committee, other committees as needed
  - **Linux Foundation** provides backend operational support, guidance on operational practices from similar projects, and program management resources to help the R Consortium achieve maximum impact.
- See https://www.r-consortium.org

Oracle Advanced Analytics technologies

Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps

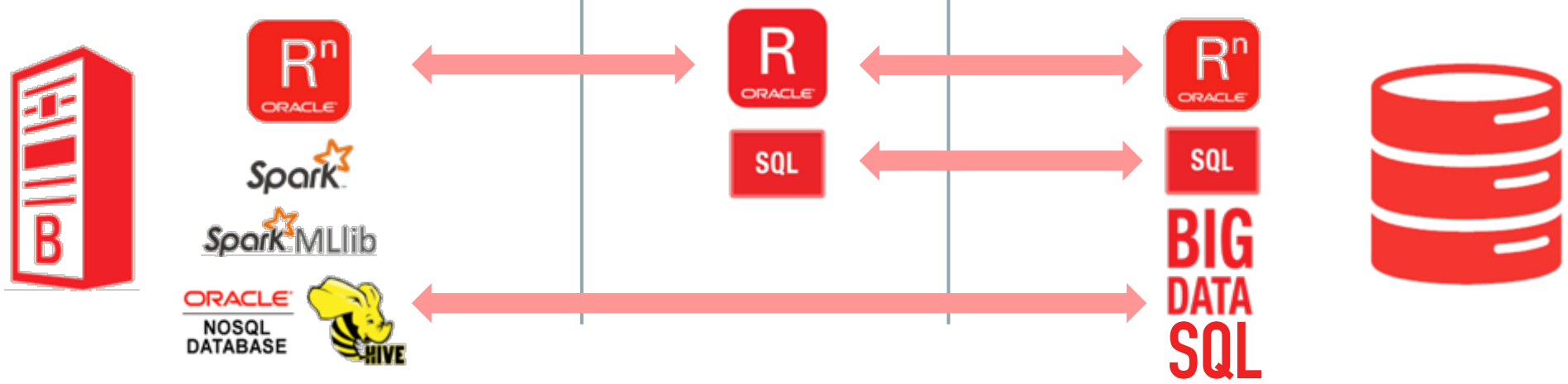
Big Data Cluster

- *ML in Spark + Spark MLlib*
- *R language transparent interface to HIVE (HQL) for data processing*
- *R platform for open-source R packages*

Client Interfaces

Oracle Database

- *ML in-Database and R Enterprise*
- *R language transparent interface to Oracle SQL for data processing*
- *R platform for open-source R packages*



Oracle Advanced Analytics technologies

Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps

Big Data Cluster

- *ML in Spark + Spark MLlib*
- *R language transparent interface to HIVE (HQL) for data processing*

Oracle R Advanced Analytics for Hadoop - Big Data Connectors license

Client Interfaces

Any SQL Client

SQL Developer

Oracle R Distribution Client

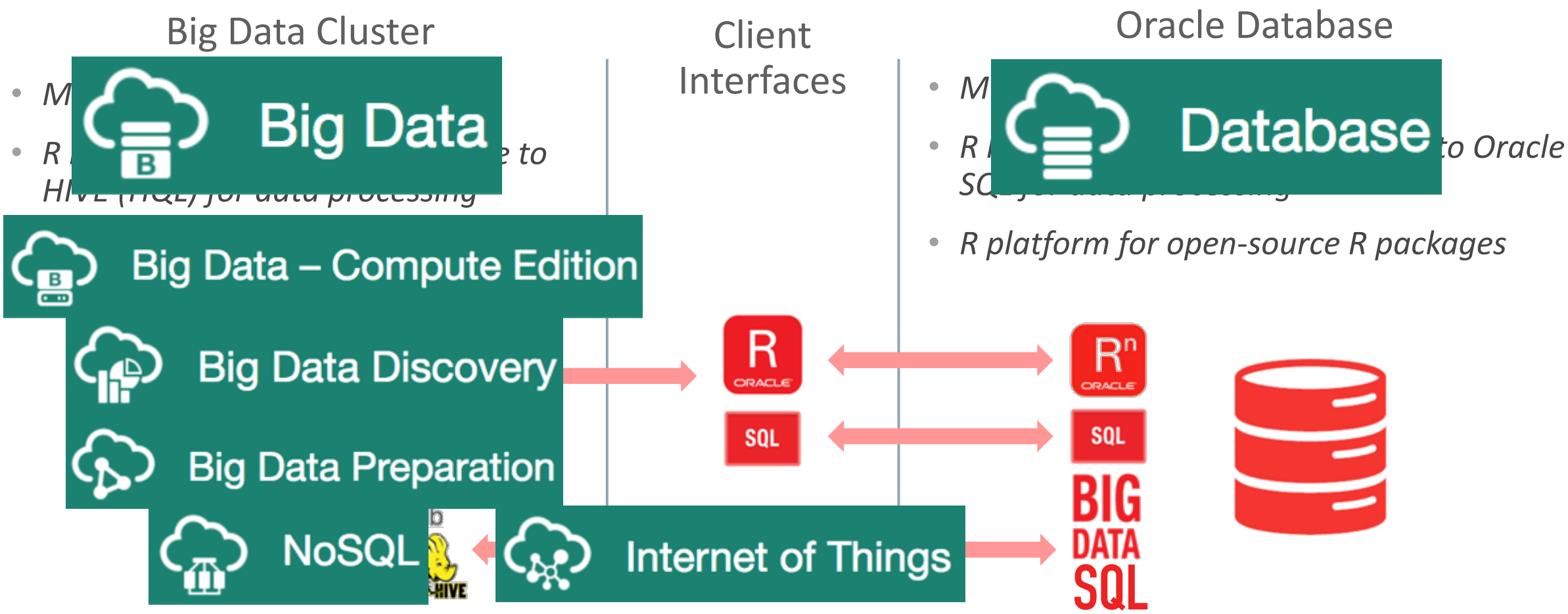
Oracle Database

- *ML in-Database and R Enterprise*
- *R language transparent interface to Oracle SQL for data processing*
- *R platform for open-source R packages*

Oracle Advanced Analytics option for the Oracle Database

Oracle Advanced Analytics technologies

Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps



Oracle Advanced Analytics

Machine Learning algorithms in-Database through SQL and Oracle R Enterprise

Classification
Logistic Regression
Decision Trees
Random Forests
Naïve Bayes
Support Vector Machines

Clustering
Hierarchical k-Means
Hierarchical O-Cluster
Expectation-Maximization

Regression
Linear Regression
Stepwise Linear Regression
Generalized Linear Models
Support Vector Machines
Multi-Layer Neural Networks

Anomaly Detection
One-Class SVM

Association Rules
Apriori

Text Mining
Tokenization
Theme Extraction

Attribute Importance
Minimum Description Length
Principal Components Analysis

Feature Extraction
Nonnegative Matrix Fact(NMF)
Singular Value Decomposition(SVD)

9,000+ Open-Source R packages

Oracle Machine Learning Algorithms

All Spark-based Algorithms by Oracle and Spark MLlib, plus HIVE and MR ones

Classification

Decision Tree (Spark MLlib)
Logistic Regression (Spark MLlib)
Logistic Regression (Oracle's Spark-based)
Support Vector Machine (SVM) (Spark MLlib)
Random Forest (Spark MLlib)
Multi-Layer Neural Networks (Oracle's Spark)

Regression

Support Vector Machine (SVM) (Spark MLlib)
Linear Model (Spark MLlib)
Linear Model (Oracle's Spark-based)
Multi-Layer Neural Networks (Oracle's Spark)
LASSO (Spark MLlib)
Ridge Regression (Spark MLlib)
Random Forest (Spark MLlib)

Clustering

Hierarchical k-Means (Spark MLlib)
Gaussian Mixture Models (Spark MLlib)

Feature Extraction & Creation

Nonnegative Matrix Factorization (Map-Reduce)
Principal Component Analysis (Spark MLlib)
Low Rank Matrix Factorization (Map-Reduce)

Transparency Functions with HIVE

Aggregations, Table Joins, summarization
Variable Creation, Push&Pull data from HIVE

Open Source R Algorithms

Ability to run any R package via our hadoop.run function in Map-Reduce mode

ORAAH Benefits: Spark-based algorithm support for R users

Efficient Native Spark-based algorithms	ORAAH	SparkR	Microsoft R Server*
Linear Regression	✓		
Generalized Linear Model	✓		
Deep Neural Networks	✓		
Spark MLlib supported Algorithms			
Linear Regression	✓	✓	✓
Generalized Linear Model	✓	✓	✓
K-Means Clustering	✓	✓	✓
Gaussian Mixture Model	✓		
LASSO (least absolute shrinkage and selection operator)	✓		
Ridge Regression	✓		
Decision Trees	✓		✓
Random Forest	✓		✓
Support Vector Machines	✓		
Principal Component Analysis	✓		
Accelerated Failure Time (AFT)		✓	
Naive Bayes Model		✓	
Stepwise Regression	Roadmap		✓

* Not executed as Spark jobs, but scale in parallel using several nodes in HPC to run proprietary engine

Source for Oracle: http://www.oracle.com/technetwork/database/database-technologies/bdc/r-advanalytics-for-hadoop/overview/index.html
Source for SparkR: https://spark.apache.org/docs/latest/sparkr.html#machine-learning
Source for Microsoft: ScaleR Datasheet at https://www.microsoft.com/en/server-cloud/products/r-server/default.aspx

ORAAH Benefits: Contrast to Microsoft R Server

Feature	ORAAH	Microsoft R on Spark
Computing Engine for Machine Learning	Native Efficient Spark-based algorithms plus Spark MLlib	Custom HPC Engine that requires proprietary input data format
Performance gains: single-threaded open-source R algorithms vs 5-node Spark Cluster	Expected at 175x on Logistic Regression	Microsoft mentions 122x on Logistic Regression
Performance gains of Spark-based algorithms processing over Map-Reduce	Observed 100-200x depending on the algorithm, which are native Java jobs in Spark	Microsoft mentions only 6x gains; this is likely only in I/O since they need to use the same proprietary algorithm for compute, outside Spark
Data Management Requirements	Data is kept in its native format, HDFS or HIVE, and loaded to memory by Spark when needed	Data needs to be converted to proprietary XdF Data format before the ScaleR algorithms can be executed

Source for Microsoft: https://channel9.msdn.com/blogs/MicrosoftR/R-Server-on-Spark
Source for Oracle: https://blogs.oracle.com/R/entry/oracle_r_advanced_analytics_for

ORAAH Benefits: Making Spark MLlib better for R users

ORAAH Formula parser can handle the full set of open-source R formula transformations, so it can be used with any Spark MLlib algorithm supported by ORAAH.

Even the current release 2.1.0 of SparkR (August '16) fails to process a simple interaction between attributes.

Using SparkMLlib Logistic Regression model in **SparkR**

```
R> model <- glm( Kyphosis ~ (Age + Number)^2, df, family = "binomial")
ERROR RBackendHandler: fitRModelFormula on org.apache.spark.ml.api.r.SparkRWrappers failed
Error in invokeJava(isStatic = TRUE, className, methodName, ...) :java.lang.IllegalArgumentException:
Could not parse formula: Kyphosis ~ (Age + Number)^2
```

Using Spark MLlib Logistic Regression model in **ORAAH**

```
R> model <- orch.ml.logistic( Kyphosis ~ (Age + Number)^2, data = data)
OBX Model Matrix: processed 1 factor variables, 0.050 sec
OBX Model Matrix: created MLlib LabeledPoint RDD (81 rows) 0.008 sec
OBX Machine Learning: MLlib Logistic Regression elapsed time 0.858 sec
R> model$coefficients
[1]  -6.568918  0.027176503  1.022537535 -0.004490547
```

Using open-source R with the same complex formula to ensure ORAAH's model coefficients are correct

```
glm( Kyphosis ~ (Age + Number)^2, data = kyphosis, family = "binomial")$coefficients
 (Intercept)          Age       Number   Age:Number
-6.568917860  0.027176503  1.022537536 -0.004490547
```

ORAAH vs Python: Spark MLlib Random Forest from HIVE

Python user steps – 47 lines

```
# Load the required Libraries
from pyspark.ml import Pipeline
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.feature import IndexToString, StringIndexer, VectorIndexer
from pyspark.sql import SparkSession
from pyspark.ml.feature import RFormula

# Create a Spark Session
if __name__ == "__main__":
    spark = SparkSession\
        .builder\
        .appName("RandomForestClassifierExample")\
        .getOrCreate()

# Establish a formula in R-style
# In Spark 2.0.0 is limited to '~', '.', ':', '+', and '-'
    formula = RFormula(
        formula="CANCELLED ~ . - FLIGHTNUM",
        featuresCol="features",
        labelCol="label")

# Load data from HIVE into Spark DataFrame
df = spark.sql("FROM ONTIME SELECT *").collect()

# Parses the input Data Frame with the Formula
dfParsed = formula.fit(df).transform(df)

# Index labels, adding metadata to the label column.
# Fit on whole dataset to include all labels in index.
labelIndexer = StringIndexer(inputCol="label", outputCol="indexedLabel").fit(dfParsed)

# Automatically identify categorical features, and index them.
# Set maxCategories so features with > 4 distinct values are treated as continuous.
featureIndexer =\
    VectorIndexer(inputCol="features", outputCol="indexedFeatures", maxCategories=4).fit(dfParsed)

# Train a RandomForest model.
rf = RandomForestClassifier(labelCol="indexedLabel", featuresCol="indexedFeatures", numTrees=50)

# Chain indexers and forest in a Pipeline
pipeline = Pipeline(stages=[labelIndexer, featureIndexer, rf])

# Train model.  This also runs the indexers.
model = pipeline.fit(dfParsed)

# Make predictions.
predictions = model.transform(dfParsed)
```

Load Libraries

Establish Spark Session

Process Formula

Copy data from HIVE

Create 3rd copy of Data for vectors

Build Model

Single Vector of Predictions

http://spark.apache.org/docs/latest/sql-programming-guide.html#hive-tables
https://github.com/apache/spark/blob/master/examples/src/main/python/ml/random_forest_classifier_example.py
https://github.com/apache/spark/blob/master/examples/src/main/python/ml/rformula_example.py

ORAAH user steps – 14 lines

```
# Load ORAAH librarires
library(ORCH)

# Connect to HIVE and Create Spark Session
ore.connect(type='HIVE',server='myhost',user='me',passwor
spark.connect(master='yarn-client',memory='8GB',dfs.namenode='my_dfs_server')

# Build the model, HIVE table as input
model <- orch.ml.random.forest(CANCELLED ~ . - FLIGHT
                               ONTIME,
                               nTrees=50)

# Predict with the model with additional columns so one can merge the results
pred <- predict(model, newdata=ONTIME, supplemental.cols="UNIQUE_ID")
```

Load Libraries

Establish HIVE and Spark Session

Build Model directly against HIVE data with full formula support

Predictions on HIVE data exported with desired columns

http://www.oracle.com/technetwork/database/database-technologies/bdc/r-advanalytics-for-hadoop/documentation/index.html

R: ORAAH Machine Learning models in **Spark**

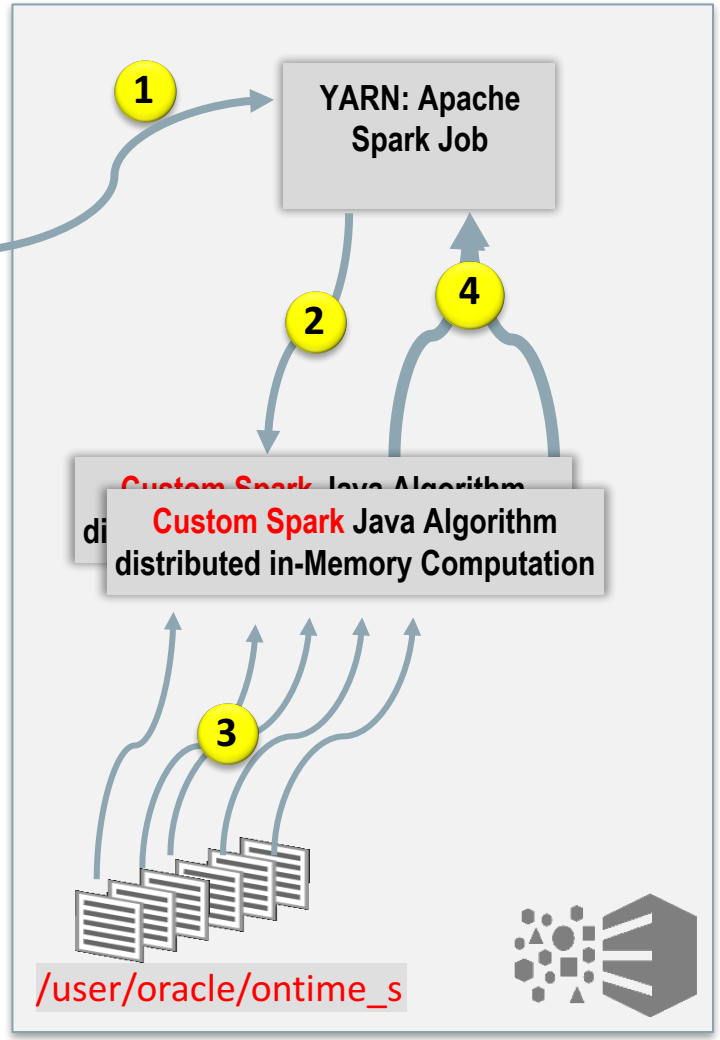
Invoke ORAAH custom parallel distributed model (GLM2 or Neural) using Spark Caching

R Console

```
Oracle Distribution of R version 3.2.0  (--) -- "Full of Ingredients"
> Connects to Spark
> spark.connect("yarn-client",memory="24g")

> # Attaches the HDFS file for use within R
> ont1bi <- hdfs.attach("/user/oracle/ontime_1bi")

> # Formula definition: Cancelled flights (0 or 1) based on other attributes
> form_oraah_glm2 <- CANCELLED ~ DISTANCE + ORIGIN + DEST + F(YEAR) + F(MONTH) +
+   F(DAYOFMONTH) + F(DAYOFWEEK)
> system.time(m_spark_glm <- orch.glm2(formula=form_oraah_glm2, ont1bi))
 ORCH GLM: processed 6 factor variables, 25.806 sec
 ORCH GLM: created model matrix, 100128 partitions, 32.871 sec
 ORCH GLM: iter  1,  deviance   1.38433414089348300E+09,  elapsed time 9.582 sec
 ORCH GLM: iter  2,  deviance   3.39315388583931150E+08,  elapsed time 9.213 sec
 ORCH GLM: iter  3,  deviance   2.06855738812683250E+08,  elapsed time 9.218 sec
 ORCH GLM: iter  4,  deviance   1.75868100359263200E+08,  elapsed time 9.104 sec
 ORCH GLM: iter  5,  deviance   1.70023181759611580E+08,  elapsed time 9.132 sec
 ORCH GLM: iter  6,  deviance   1.69476890425481350E+08,  elapsed time 9.124 sec
 ORCH GLM: iter  7,  deviance   1.69467586045954760E+08,  elapsed time 9.077 sec
 ORCH GLM: iter  8,  deviance   1.69467574351380850E+08,  elapsed time 9.164 sec
user  system elapsed
84.107   5.606 143.591
```

Oracle R Advanced Analytics for Hadoop Client Packages

Spark-Based Machine Learning algorithms module



R: Machine Learning models using **Spark MLlib** algorithms

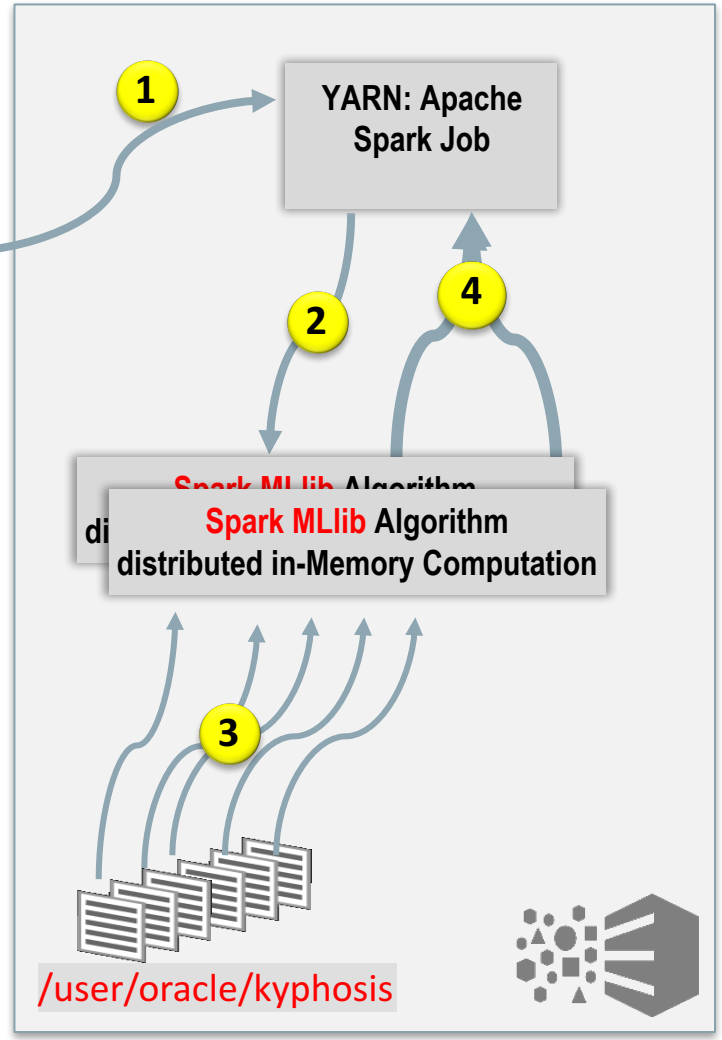
Invoke ORAAH custom interface to Spark Mllib algorithms within R

R Console

```
Oracle Distribution of R version 3.2.0  (--) -- "Full of Ingredients"
> Connects to Spark
> spark.connect("yarn-client",memory="24g")

> # Attaches the HDFS file for use within R
> # Can use either HDFS input data or HIVE tables
> data <- hdfs.attach("/user/oracle/kyphosis")

# Building a LASSO model with Spark MLlib from a one line of code in R
> model <- orch.ml.lasso(formula = Kyphosis ~ Number + Age, data = data)
OBX Model Matrix: processed 1 factor variables, 0.155 sec
OBX Model Matrix: created MLlib LabeledPoint RDD (81 rows) 0.015 sec
OBX Machine Learning: MLlib Lasso elapsed time 3.582 sec

# Scoring a LASSO model with Spark Mllib from a one line of code in R
> pred <- predict(model, newdata = data, supplemental = c("Kyphosis", "Age"))
OBX Model Matrix: created predict RDD (81 rows) 0.007 sec
```

Oracle R Advanced Analytics for Hadoop Client Packages

Spark-Based Machine Learning algorithms module

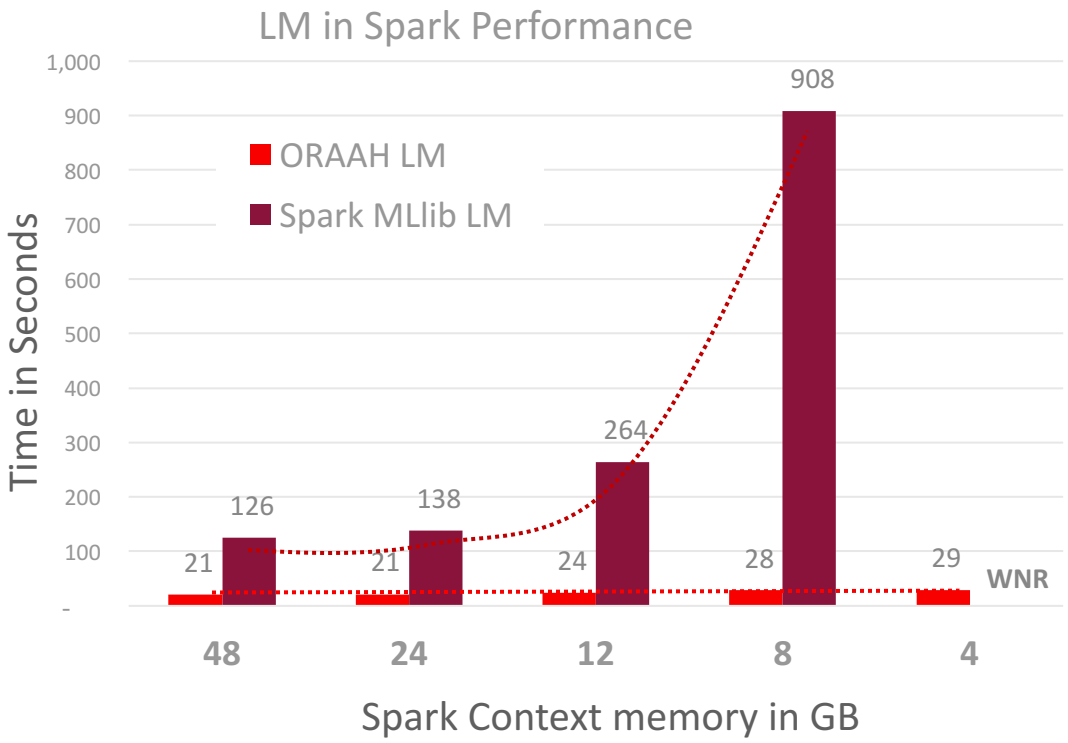


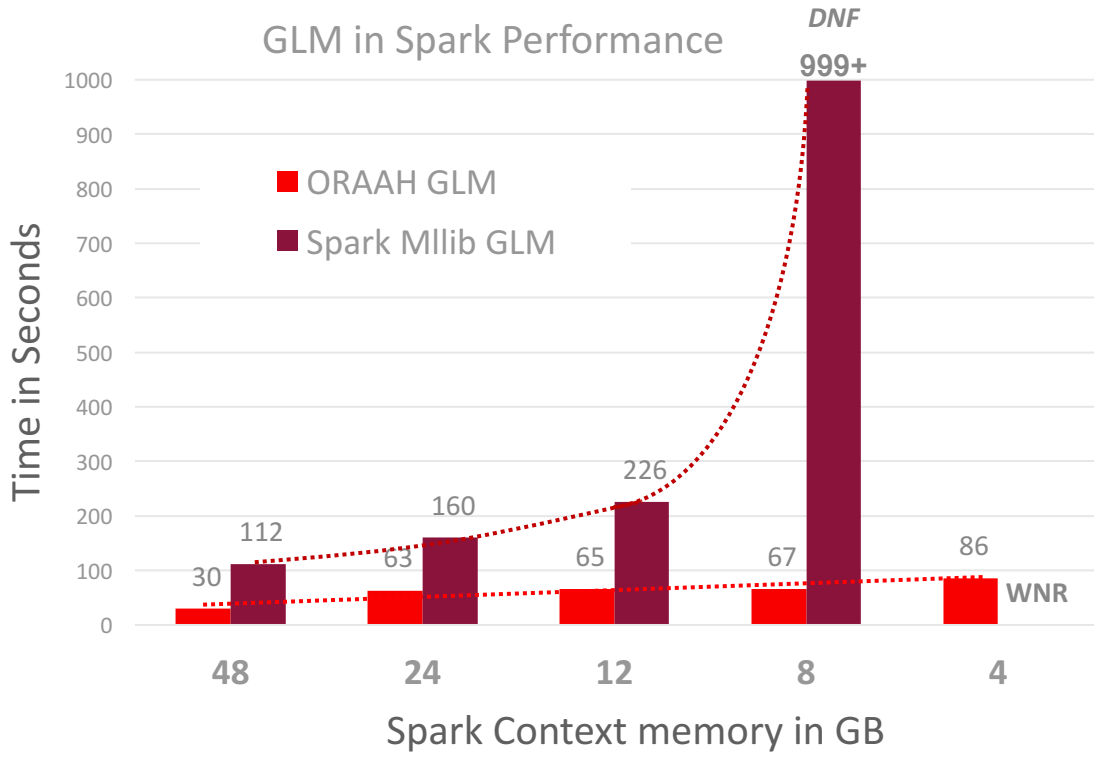
New Functionality for ORAAH 2.6.0 – Spark MLlib Interface
Build model and Score model using 1 line of R code for each

ORAAH Benefits: benchmark vs Spark MLlib

Comparing performance of LM and GLM on varying Spark memory footprints

Benchmark on single X5-2 Node with 74 threads and 256 GB of Total RAM, Spark 1.6.0 on CDH 5.8.0
Input Data is 15GB ”Ontime” airline dataset with 123mi records, predicting 8,926 total coefficients



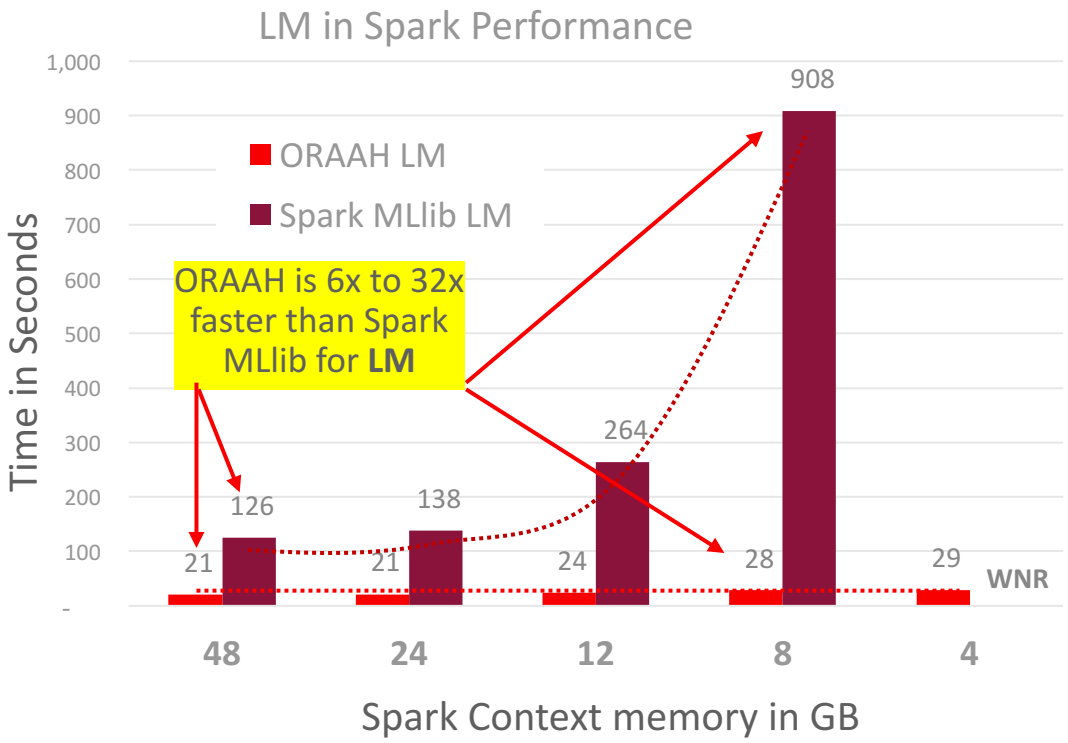


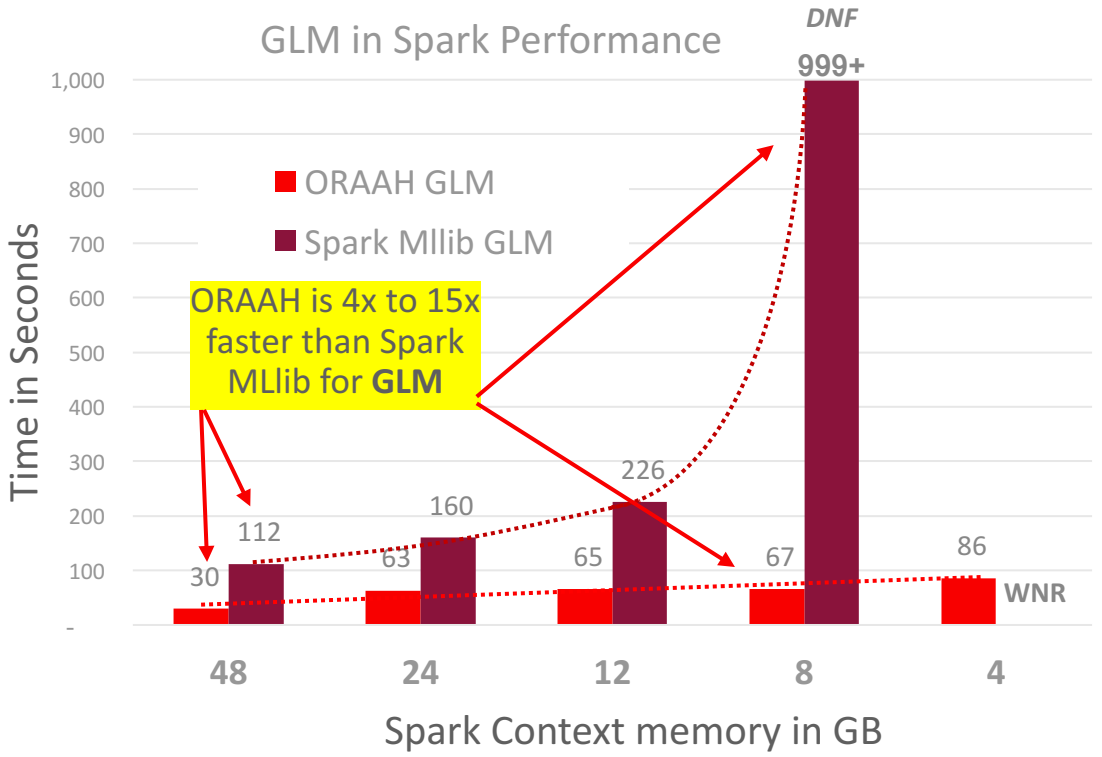
LM formula: ARRDELAY ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)
GLM formula: CANCELLED ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)

ORAAH Benefits: benchmark vs Spark MLlib

Comparing performance of LM and GLM on varying Spark memory footprints

Benchmark on single X5-2 Node with 74 threads and 256 GB of Total RAM, Spark 1.6.0 on CDH 5.8.0
Input Data is 15GB ”Ontime” airline dataset with 123mi records, predicting 8,926 total coefficients



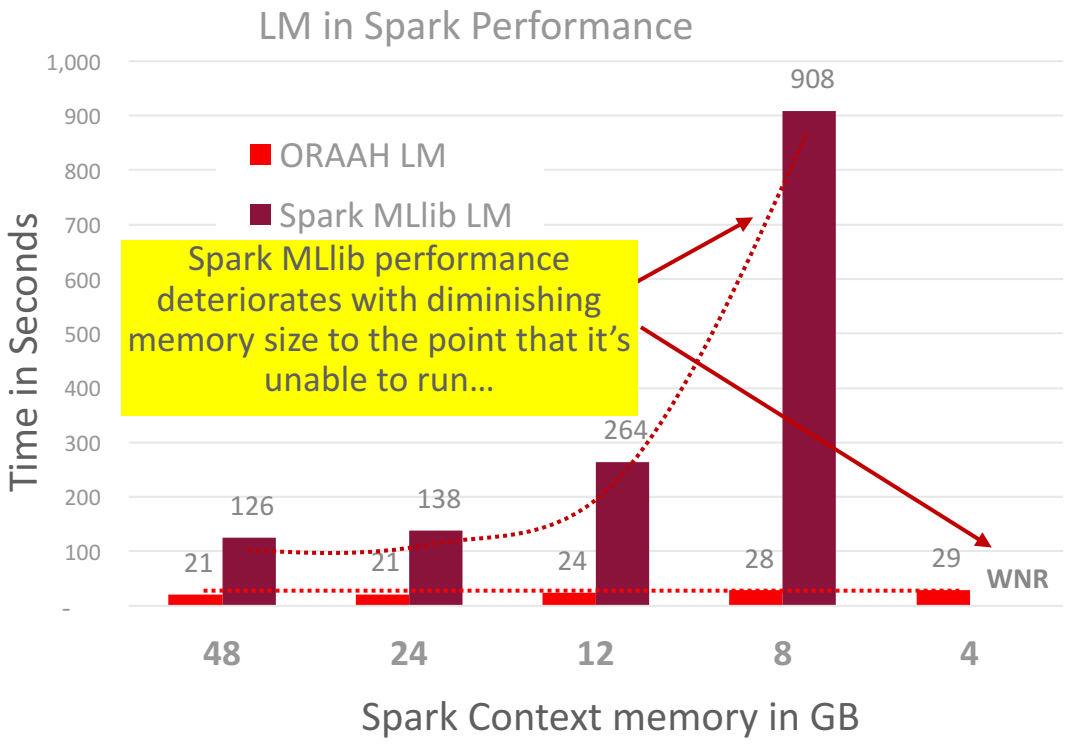


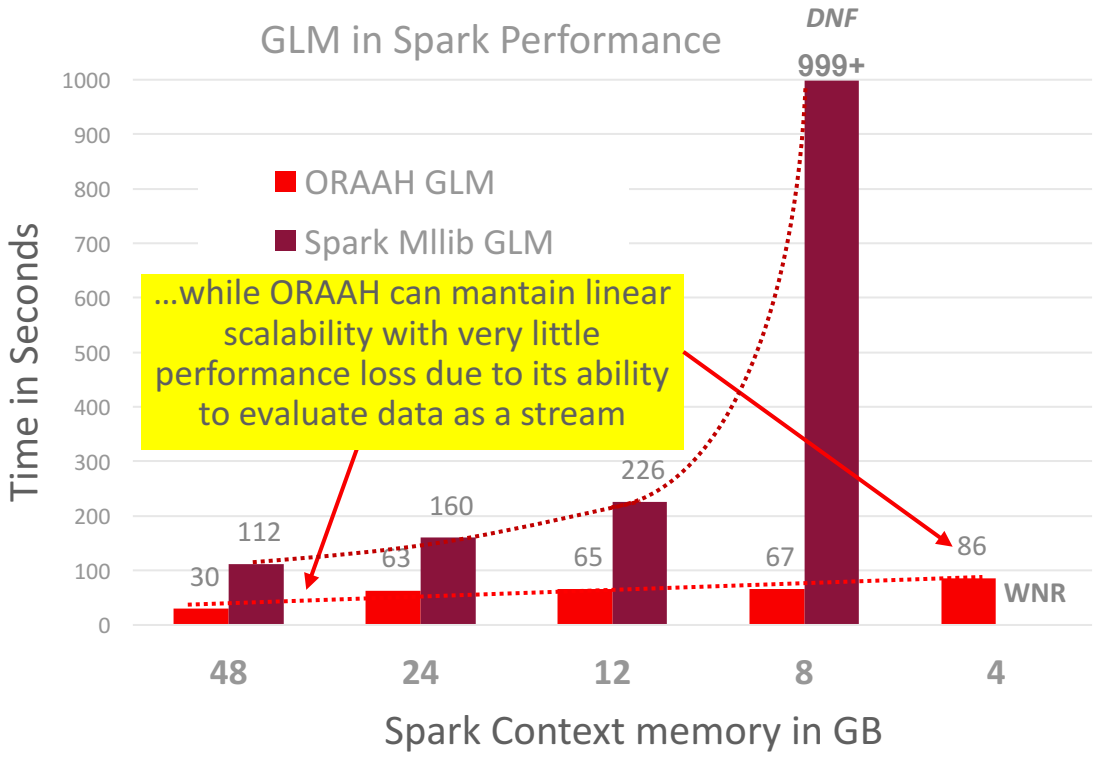
LM formula: ARRDELAY ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)
GLM formula: CANCELLED ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)

ORAAH Benefits: benchmark vs Spark MLlib

Comparing performance of LM and GLM on varying Spark memory footprints

Benchmark on single X5-2 Node with 74 threads and 256 GB of Total RAM, Spark 1.6.0 on CDH 5.8.0
Input Data is 15GB ”Ontime” airline dataset with 123mi records, predicting 8,926 total coefficients

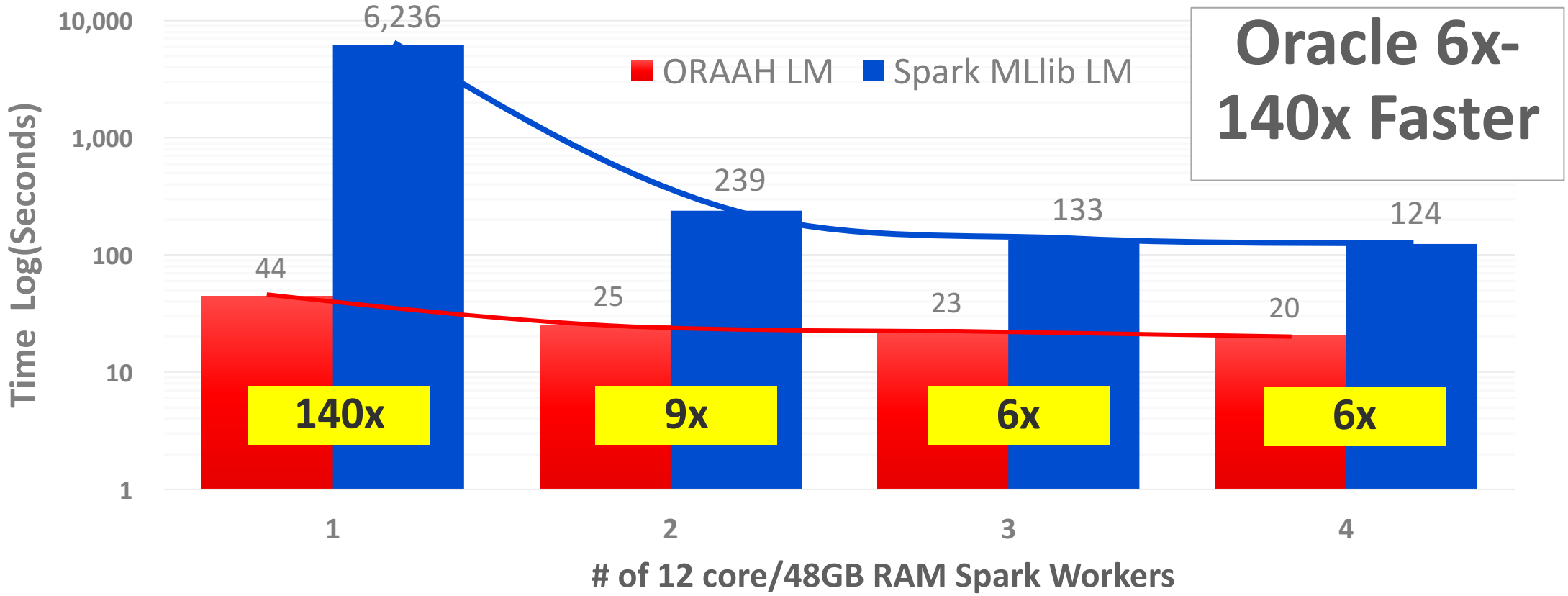




LM formula: ARRDELAY ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)
GLM formula: CANCELLED ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)

Faster Time to Results: Machine Learning

Oracle ML outperforms Spark MLlib



Linear Model algorithm over airline data, Predict 8,000+ coefficients over 120M records

Getting started

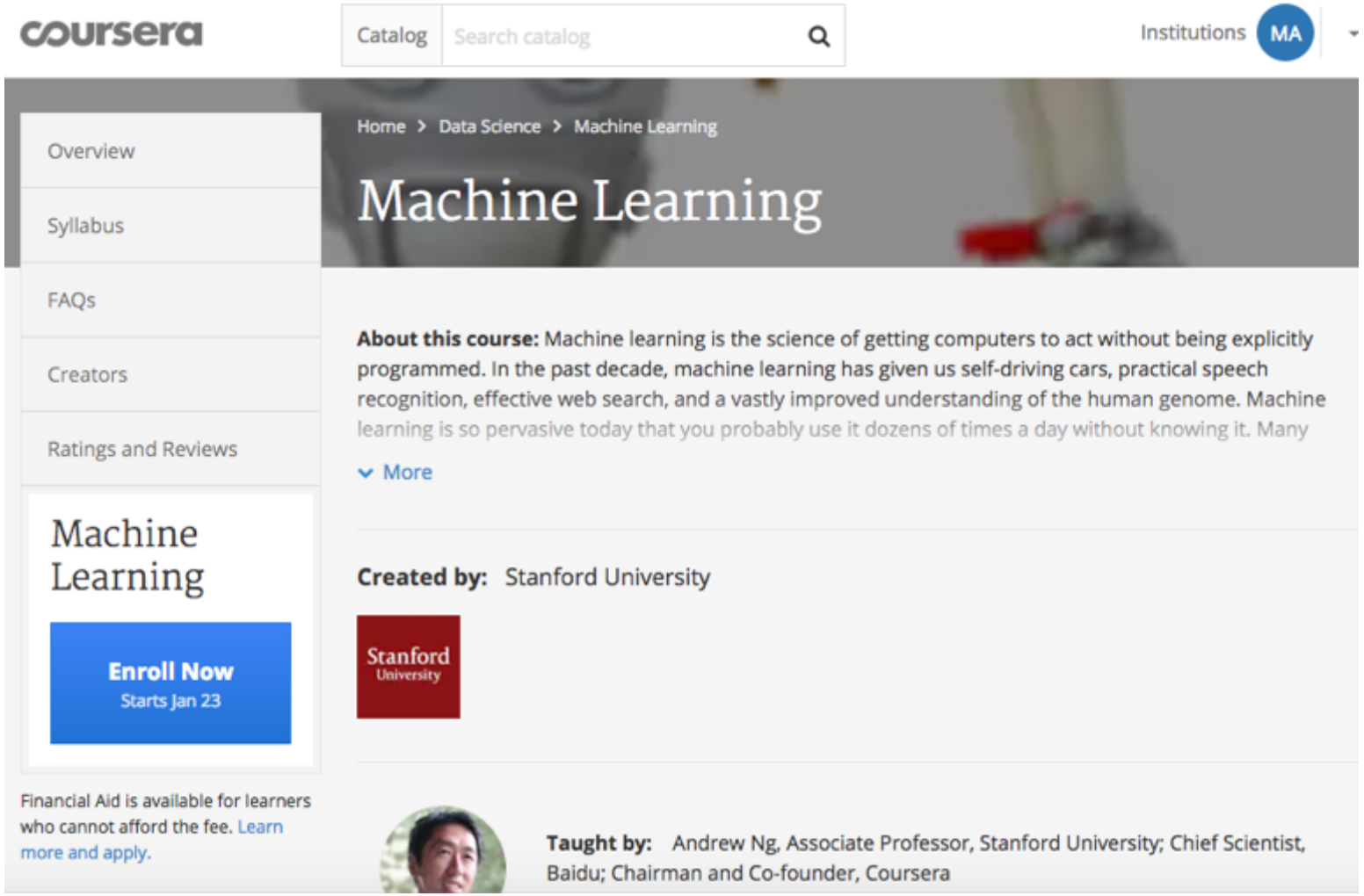
Where to learn the Concepts of Machine Learning?

ml-class.org
coursera.org

By Andrew Ng
Associate Professor, Stanford University;

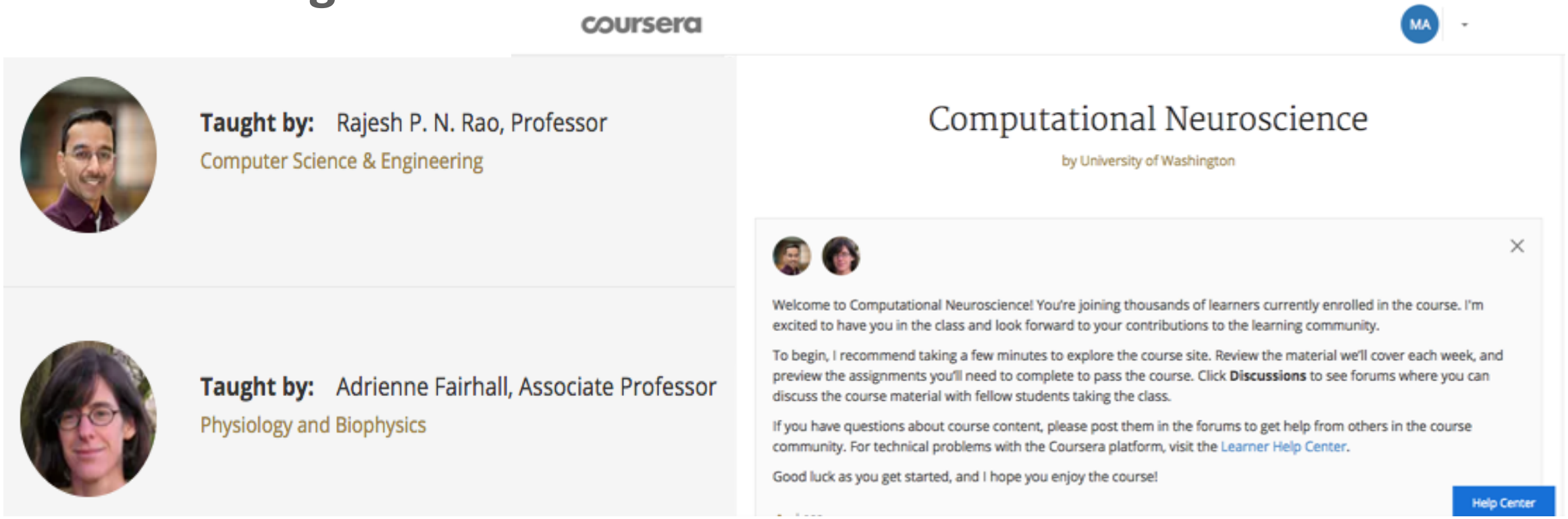
Chief Scientist, Baidu;

Chairman and Co-founder, Coursera



Where to learn the Concepts of Neuroscience?

coursera.org



https://www.coursera.org/learn/computational-neuroscience/lecture/iynBe/1-4-the-electrical-personality-of-neurons

Getting started: OAA Links and Resources

Oracle Advanced Analytics Overview:

- Big Data Analytics with Oracle Advanced Analytics: Making Big Data and Analytics Simple white paper on OTN
- Oracle Advanced Analytics Customer Successes: Data Mining Customers
- Oracle Advanced Analytics Customer Successes: Oracle R Customers

YouTube recorded OAA Presentations and Demos:

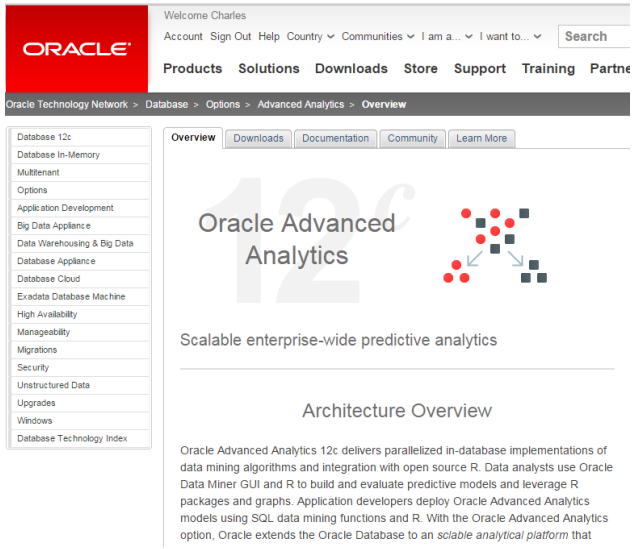
- Oracle Advanced Analytics and Data Mining ("live" Demos on ODM'r 4.0 New Features, Retail, Fraud, Loyalty, Overview, etc.)
- Oracle Advanced Analytics Youtube Channel: Data Mining and R

Getting Started:

- Link to OAA/Oracle Data Miner Workflow GUI Online (free) Tutorial Series on OTN
- Link to OAA/Oracle R Enterprise (free) Tutorial Series on OTN
- Link to Free Oracle Advanced Analytics **"Test Drives" on Oracle Cloud** via Vlamis Partner
- Link to Getting Started w/ ODM blog entry
- Link to New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course.
- Link to OAA/Oracle R Enterprise 2-Day Instructor-led Oracle University course
- Oracle Data Mining Sample Code Examples

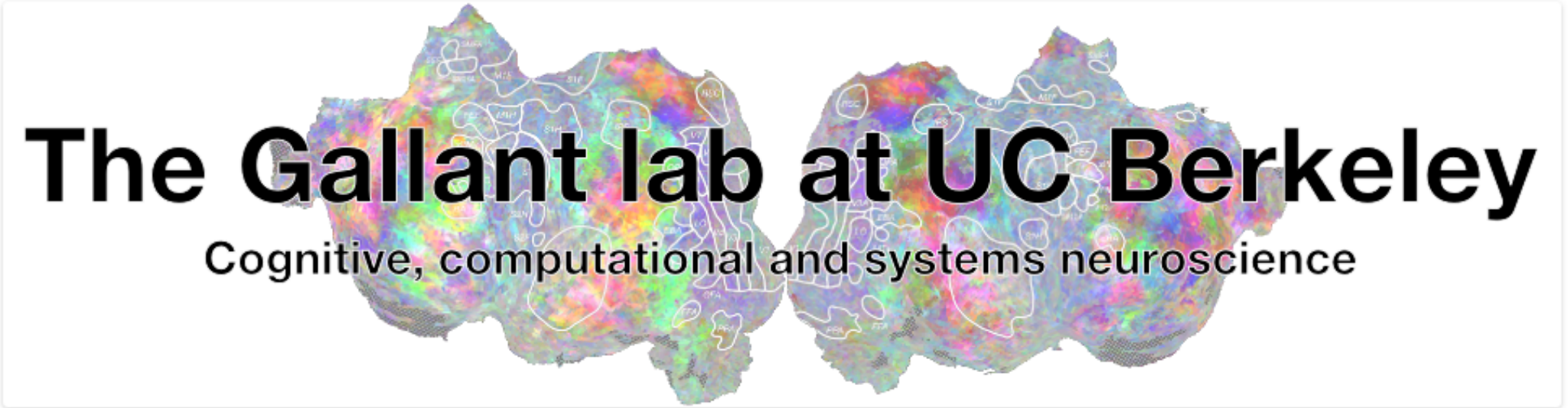
Additional Resources:

- Oracle Advanced Analytics Option on OTN page
- OAA/Oracle Data Mining on OTN page, ODM Documentation & ODM Blog
- OAA/Oracle R Enterprise page on OTN page, ORE Documentation & ORE Blog
- Oracle R Advanced Analytics for Hadoop (ORAAH) on OTN
- **Business Intelligence, Warehousing & Analytics—BIWA Summit'17**, Jan 31, Feb 1 & 2, 2017 at Oracle HQ Conference Center (w/ links to customer presentations)



http://gallantlab.org/index.php/brain-viewer/

NEWS **BRAIN VIEWER** PUBLICATIONS PEOPLE PRESS RESEARCH LINKS JOIN



Brain viewer

Our laboratory has developed powerful visualization tools that provide an easy way to interact with the data online. (The open source software, Py-cortex, can be obtained from here.) This page provides links to several online viewers, customized for specific data sets.


ORACLE®