



KNOWLEDGE MANAGEMENT

Knowledge Discovery

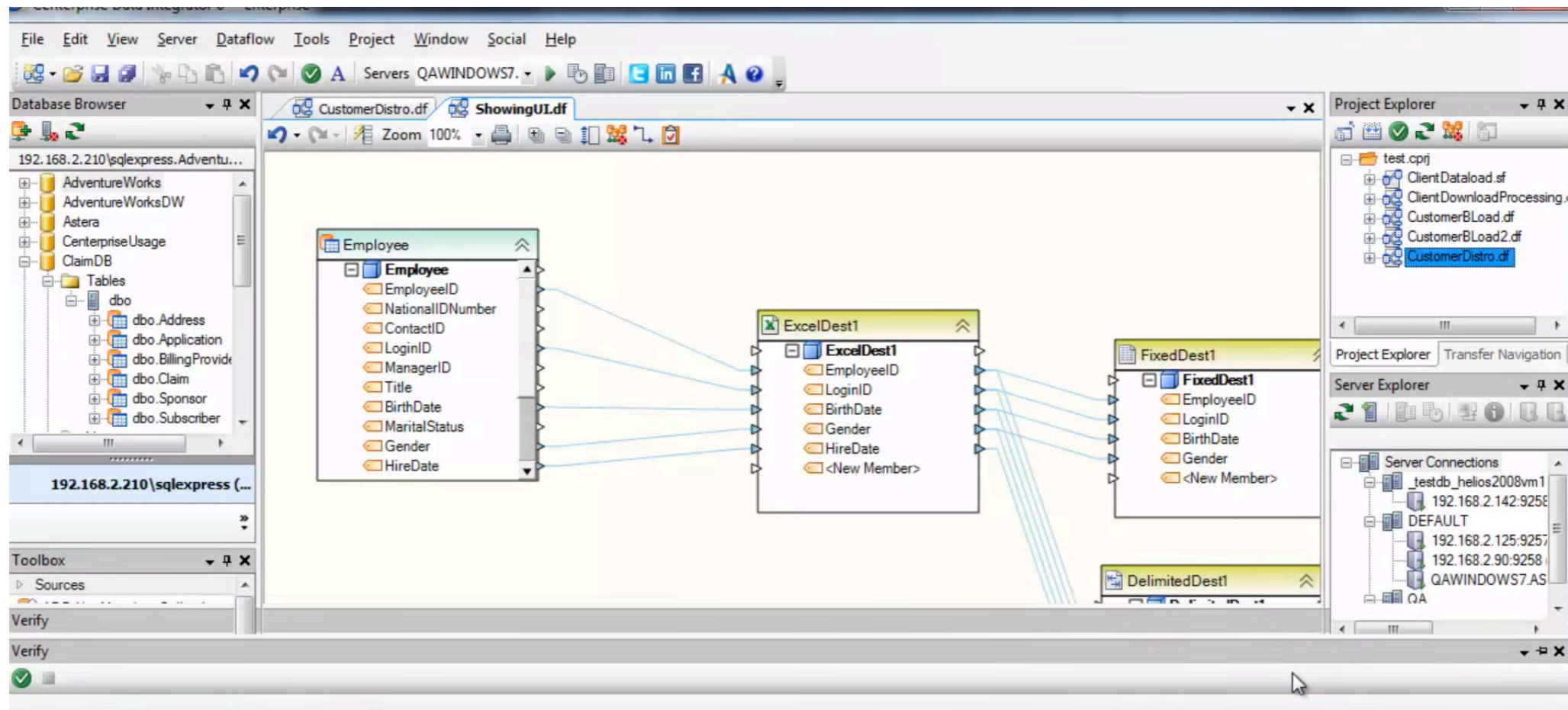


DATA MINING TOOL

- Data mining is looking for **hidden, valid, and all the possible useful patterns** in large size data sets
- Data Mining is a technique which helps you to **discover** unsuspected/undiscovered **relationships** amongst the data for business gain
- According to Guru99, There, are many useful tools available for Data mining
- Following is a curated list of Top 10 handpicked **Data Mining software** with popular features

1). REPORT MINER

- An enterprise-grade data mining solution that enables business users to **create reusable extraction** templates in a drag-and-drop user interface
- Features:
 - Extract data from a range of unstructured sources, including PDFs, TXT, DOC, DOCX, and more
 - Create reusable extraction templates to mine data from documents containing similar layout
 - Automate data mining process with features like job scheduling, email/folder/FTP integration, automated address and name parsing, and use event-based triggers to run workflows
 - Use built-in data quality and profiling transformations to validate data



The screenshot displays the SQL Server Data Tools (SSDT) interface. The main workspace shows a data flow diagram for a dataflow task named 'CustomerDistro.df'. The source is an 'Employee' table with columns: EmployeeID, NationalIDNumber, ContactID, LoginID, ManagerID, Title, BirthDate, MaritalStatus, Gender, and HireDate. The data is mapped to two destinations: 'ExcelDest1' and 'FixedDest1'. 'ExcelDest1' receives EmployeeID, LoginID, BirthDate, Gender, and HireDate. 'FixedDest1' receives EmployeeID, LoginID, BirthDate, and Gender. A third destination, 'DelimitedDest1', is present but has no data flow connections.

The bottom pane shows the following error messages:

Severity	Name	Context	Message
Error	DelimitedDest1		Destination file path must be provided.
Error	ExcelDest1		Destination file path must be provided.
Error	FixedDest1	EmployeeID	Invalid length <0>.
Error	FixedDest1	LoginID	Invalid length <0>.
Error	FixedDest1	BirthDate	Invalid length <0>.

2). OCTOPARSE

- SaaS web data platform that satisfies users' most crawling needs, both basic and advanced. You can use it to scrape web data and turn unstructured/semi-structured data into structured data sets without coding
- **Features:**
 - Two kinds of operation mode - Wizard Mode and Advanced Mode - for non-programmers to quickly pick up
 - User-friendly point-and-click interface
 - Provides Scheduled Cloud Extraction to extract dynamic data in real-time and keeps track of any website updates
 - Uses the built-in RegEx tools and XPath configuration to locate elements precisely on complex websites
 - Offers IP Proxy Servers that automate the IPs, greatly reducing the chances of being detected by aggressive websites



Octoparse Version 6.0

Task: yelp

1 Set Basic Information 2 Design Workflow 3 Extraction Options 4 Done

Workflow Designer

```
graph TD; A[Go to the Webpage] --> B[Enter Text]; B --> C[Click Item]; C --> D[Cycle Pages]; D --> E[Loop Item]; E --> F[Click Item]; F --> G[Extract Data]; G --> H[Loop Item]; H --> D;
```

Customize Current Action

PageUrl: http://www.yelp.co.uk/

Advanced Options

Action Caption: Go to the Webpage Timeout: 120 seconds

Block Pop-up: Block popup windows (possibly Ads)

Use Loop URL: Use the current loop items as navigation URLs

Scroll Down: Scroll down to page bottom when finished loading

Cache Settings

Retry in following conditions:

Irvine Restaurants, Dentists, Pubs, Beauty Salons, Doctors | Data Schema | Work Flow | Help

https://www.yelp.co.uk/irvine-ca-us

Block Pop-up

yelp Find pizza, pub, Fox & Hound Near Irvine, CA Sign Up Log In

Home About Me Write a Review Find Friends Messages Talk Events

Cookies help us deliver our services. By using our services, you agree to our use of cookies. Learn more. OK

Yelo Irvine London Edinburgh Glasgow Manchester Leeds Belfast

Welcome fangsheng, You are professional account. The expiration date is 04/15/2017

Element :DIV

3). SAS DATA MINING

- Statistical Analysis System is a product of SAS. It was developed for **analytics and data management**. It offers a **graphical UI for non-technical users**
- **Features:**
 - SAS Data mining tools help you to analyze Big data
 - It is an ideal tool for Data mining, text mining & optimization
 - SAS offers distributed memory processing architecture which is highly scalable

SAS Visual Analytics - *Both VS VDMML 8.1 Partition Examples

Logistic Regn Partition Forest Partition Gradient Boost Partition **Neural Net Partition** Model Comparison

Search SAS Test User 1

Objects

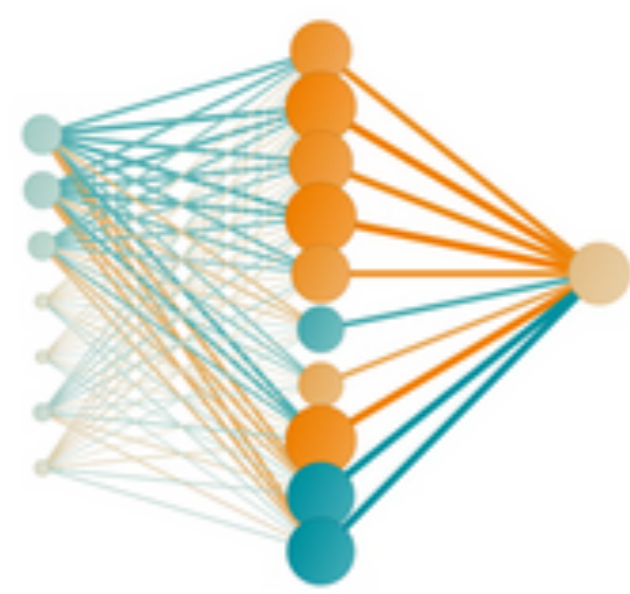
- Slider
- Text Input
- Analytics**
 - Network Analysis
 - Text Topics
- Other**
 - Container
 - Image
 - Prompt Container
 - Text
 - Web Content
- SAS Visual Statistics**
 - Cluster
 - Decision Tree
 - Generalized Linear Model
 - Linear Regression
 - Logistic Regression
 - Model Comparison
- SAS Visual Data Mining and Machine Learning**
 - Factorization Machine
 - Forest
 - Gradient Boosting
 - Neural Network
 - Support Vector Machine

Drop a data item or control to create a report prompt

Drop a data item or control to create a page prompt

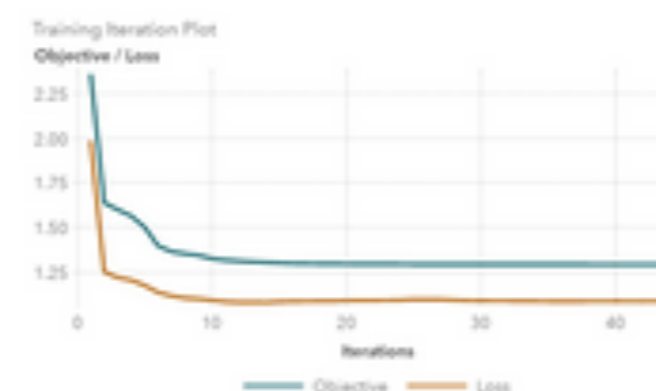
Neural Network **BAD (CAT)** (event=1) Validation Misclassification **0.0771** Observations Used **3,743** Unused **2,217**

Network

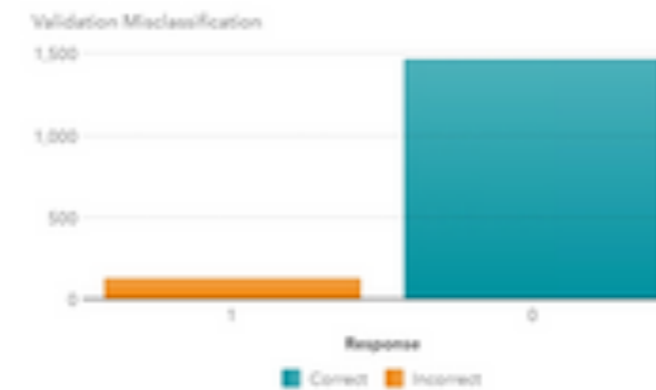


Training Iteration Plot

Objective / Loss



Validation Misclassification



Weight

Neuron Absolute Average: 0.5309, 0.0591

Neuron Average: -0.50 to 0.53

Link Absolute: 0.5309, 0.0607

Link: -0.50 to 0.53

Response: Correct (blue), Incorrect (orange)

Options

- General
- Background
- Neural Network**
 - Use validation partition
 - Include missing
 - Standardization: Midrange
 - Use default number of iterations: 250
 - Use default maximum time(sec): 270
 - Optimization method: LBFGS
 - L1: 0
 - L2: 0.1
 - Hidden layers: 1
 - Allow direct connections between input and target neurons
 - Hidden layer E:
 - Neurons: 10

4). TERADATA

- It is a massively parallel open processing system for developing **large-scale data warehousing** applications
- Run on **Unix/Linux/Windows** server platform
- Features:
 - Teradata Optimizer can handle up to 64 joins in a query
 - Tera data has a low total cost of ownership. It is easy to set up, maintain, and administrate
 - It supports SQL to interact with the data stored in tables. It provides its extension
 - It helps you to distribute the data to the disks automatically with no manual intervention
 - Teradata provides load & unload utilities to move data into/from Teradata System.

Aqua Data Studio 15.0.0 [Untitled]*

File Edit Server Query Automate Query Builder Visual Analytics ER Modeler Tools DBA Tools Window Help

SQL Server 2000 .54 bi

Servers

- Teradata 14.0
 - My Database
 - Databases
 - All
 - Crashdumps
 - DataType
 - DBC
 - dbtest1
 - dbtest2
 - dbtest4
 - Default
 - fcy_test_user
 - Tables
 - allan_
 - atest
 - Categ
 - Categ
 - CHILD
 - Custor
 - Custor
 - Custor
 - Custor
 - Custor
 - Custor
 - Custor
 - Custor
 - Emplo
 - Emplo
 - Emplo

DBC@Teradata 14.0 [Untitled]*

Database: fcy_test_user_1 Schema: DBC Username: DBC sid: 1060

```

1 select * from Orders
2 go
  
```

2 | 3 : 0 | INS PC | [5/12/2014 12:52:09 PM] Script executed - No Errors [Time: 3s]

Grid Pivot Grid Form Execution Plan Text Text History Client Statistics

830 record(s) [Fetch MetaData: 1ms] [Fetch Data: 380ms]

Freight

Drop Data Fields Here ShipVia ↑

ShipCountry	1	2	3
Argentina	5	7	4
Austria	12	15	13
Belgium	3	8	8
Brazil	31	35	17
Canada	4	10	16
Denmark	6	5	7
Finland	8	5	9
France	27	29	21
Germany	41	53	28
Ireland	4	9	6
Italy	14	9	5

Field List (Drag Items to the Pivot Grid):

- CustomerID
- EmployeeID
- OrderDate
- OrderID
- RequiredDate
- ShipAddress
- ShipCity
- ShipName
- ShipPostalCode
- ShipRegion
- ShippedDate

Add To Row Area

Local Database Servers / NextGen Servers / Teradata / Teradata 14.0 / Databases / fcy_test_user_1 / Tables / Orders

165 : 202 : 2,513 MB

5). R-PROGRAMMING

- R is a language for statistical computing and graphics. It also used for **big data analysis**
- It provides a wide variety of **statistical tests**
- **Features:**
 - Effective data handling and storage facility
 - It provides a suite of operators for calculations on arrays, in particular, matrices
 - It provides a coherent, integrated collection of big data tools for data analysis
 - It provides graphical facilities for data analysis which display either on-screen or on hardcopy



R File Edit Format Workspace Packages & Data Misc Window Help 100% Tue 2:14 PM stefano iacus

R Console

```
rgl.sr> ylen <- ylim[2] - ylim[1] + 1
rgl.sr> colorlut <- terrain.colors(ylen)
rgl.sr> col <- colorlut[y - ylim[1] + 1]
rgl.sr> rgl.clear()
rgl.sr> rgl.surface(x, z, y, color = col)
```

R Data Editor

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

Quartz (2) - Active

Given : depth

R Workspace Browser

Object	Type	Structure
dati	data.frame	dim: 20 4
g	factor	levels: 10
l	numeric	length: 12
n	numeric	length: 1
opar	list	length: 2
pie.sales	numeric	length: 6
pin	numeric	length: 2
scale	numeric	length: 1
usr	numeric	length: 4
women	data.frame	dim: 15 2
height	numeric	length: 15
weight	numeric	length: 15
x	numeric	length: 87

R Package Manager

status	Package	Description
<input checked="" type="checkbox"/> loaded	graphics	The R Graphics Package
<input type="checkbox"/> not loaded	grid	The Grid Graphics Package
<input type="checkbox"/> not loaded	lattice	Lattice Graphics
<input checked="" type="checkbox"/> loaded	methods	Formal Methods and Classes
<input type="checkbox"/> not loaded	mgcv	GAMs with CCV smoothness estimation

```
BoxDens=function(data, npts = 200., x = c(0., 1.), y = c(0., 1.), add = TRUE, col = 11., border=FALSE, collin)
{
  dens <- density(data, n = npts)
  dx <- dens$x
  dy <- dens$y
  if(add == FALSE)
    plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
         ylab = "")
  if(orientation == "paysage") {
    dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x[1.])
    dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y[1.])
    seqbelow <- rep(y[1.], length(dx))
    if(Fill == T)
      confshade(dx2, seqbelow, dy2, col = col)
    if (border==TRUE) points(dx2, dy2, type = "l", col = col)
  }
  else {
    dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y[1.])
  }
}
```

RGL device 1 (active)

6). BOARD

- Board is a Management Intelligence Toolkit. It combines features of **business intelligence** and **corporate performance management**
- It is designed to deliver **business intelligence and business analytics** in a single package
- **Features:**
 - Allows you to Analyze, simulate, plan and predict using a single platform
 - To build customized analytical and planning applications.
 - Board All-In-One combines BI, Corporate Performance Management, and Business Analytics.
 - It empowers businesses to develop and maintain sophisticated analytical and planning applications
 - The proprietary platform helps to report by accessing multiple data sources



7). DUNDAS

- Dundas is an enterprise-ready Data mining tool which can be used for building and **viewing interactive dashboards, reports**, etc.
- Able to deploy **Dundas BI** as the central data portal for the organization
- **Features:**
 - Server application with full product functionality
 - Integrate and access all kind of data sources
 - Customizable data visualizations
 - Smart drag and drop tools
 - Visualize data through maps
 - Predictive and advanced data analytics

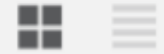
Welcome home, Jean-Luc Picard

PROJECTS > All

Search



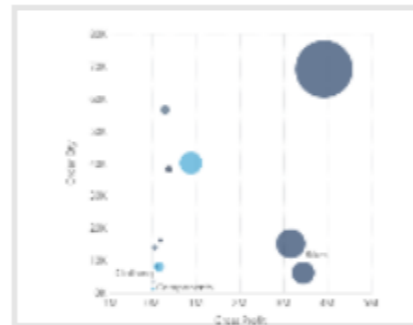
Getting Started with Dundas BI > Dashboards > Basic



- Explore Data
- New Dashboard
- New Slideshow
- New Metric Set
- New Hierarchy
- New Data Cube
- Administration

Customize Home

Basic 9 Items



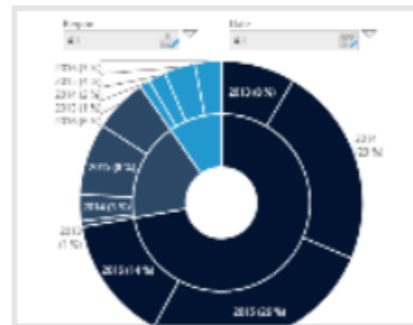
Bubble chart - Gross profit vs. orders by country and...



Line chart - Period over period



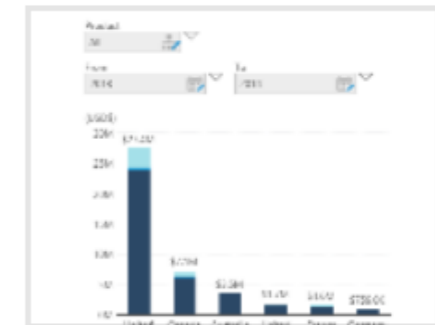
Map - Sales by country



Pie chart - Sales per year per region



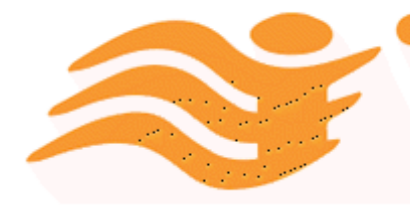
Pop-up dashboard - Order per product over months



Stacked bar chart - Sales per year per product, sort...

8). INETSOFT

- Inetsoft allows the **quick and flexible transformation** of data from various sources
- **Features:**
 - It helps you to access structured and semi-structured sources, on-premise applications
 - Allows you to optimize apps for data consumption and updating
 - Offer customized and secure levels of data exploration and reporting
 - Scale up for large data sets of users using Inbuilt Spark platform
 - Generate paginated reports with embedded business logic and parameterization



Home

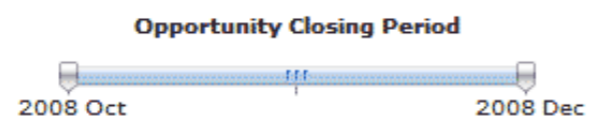
About Style Intelligence

Style Intelligence

<< Style Intelligence



Welcome salesforce.etl@inetsoft.com
Data updated: 24-Oct-2008 13:28:44



Color Points By: Probability R
Size Points By: Amount

Fluctuations From

- Week Of**
- Oct-18-2008
 - Oct-11-2008
 - Oct-04-2008
 - Sep-27-2008

- Account Type**
- - Customer - Chann
 - Customer - Direct

To Current Period
 Oct-24-2008

- Opportunity Type**
- - Existing Customer
 - Existing Customy

- Owner**
- Chris Turner

- Location**
- Australia
 - Austria
 - Belgium

- Stage**
- Id. Decision Make
 - Needs Analysis
 - Negotiation/Revie
 - Perception Analys
 - Proposal/Price Q

- Lead Source**
- - Employee Referral
 - External Referral

- Probability Range**
- 0-20
 - 20-40
 - 40-60
 - 60-80
 - 80-100

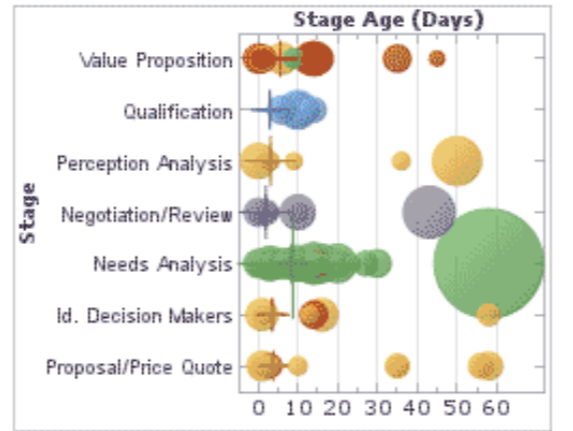
- Campaign Source**
- - Web Search 2008
 - Webinar 200809

Explore

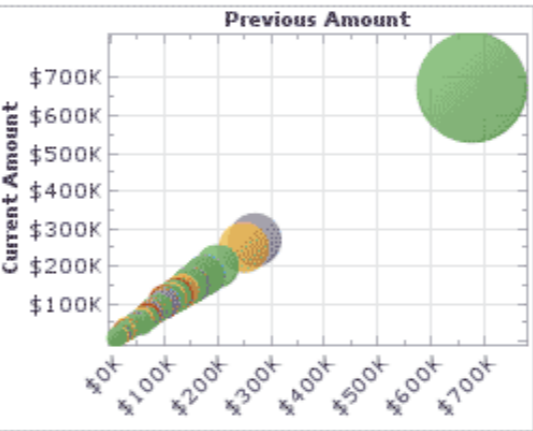
- Dashboards
 - Analysis
 - Account Analysis
 - Campaign Analysis
 - Lead Analysis
 - Pipeline Fluctuations**
 - Pipeline Trend Analysis
 - Won Lost Analysis
 - Executive Sales Dashboard
- Reports

Administration
About

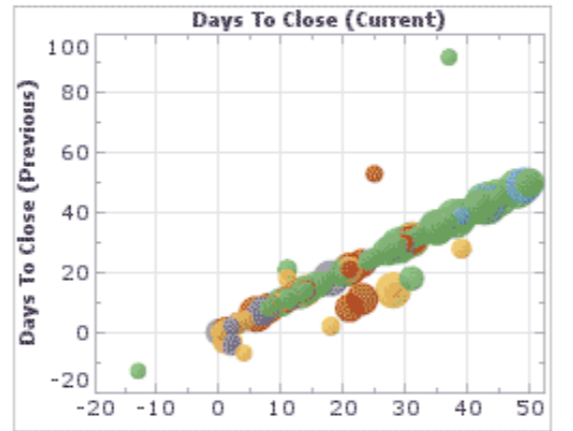
Pipeline Velocity (+Avg Age for last 12 months)



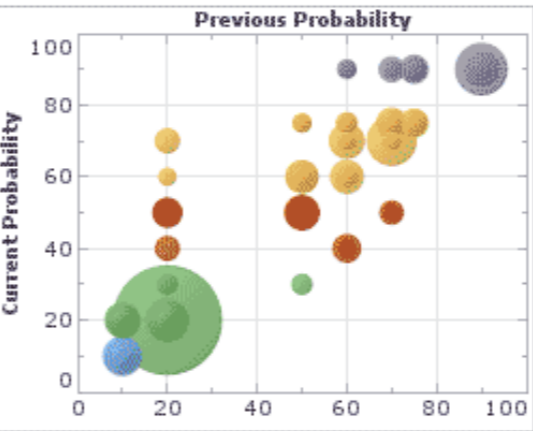
Amount Fluctuations



Closing Fluctuations



Probability Fluctuations



9). H3O

- H3O is used to perform data analysis on the data held in **cloud computing application** systems
- **Features:**
 - H3O allows you to take advantage of the computing power of distributed systems and in-memory computing
 - It allows fast and easy deployment into production with Java and binary format
 - It helps you to use the programming languages like R, Python and others to build a model in H3O
 - Distributed, In-memory Processing



H2O.ai Experiment 91e471

Show Experiments

1.0.4

TRAINING DATA

DATASET
BNPParibas-train.csv

ROWS: 114K COLUMNS: 133 DROPPED COLS: 0 TEST DATASET: Yes

TARGET COLUMN

target

TYPE	COUNT	UNIQUE	FREQ
Int	114321	2	27300

SCORED 393/426 MODELS ON 4318 FEATURES



ELAPSED: 00:14:48 ITERATION: 47/50

FINISH

EXPERIMENT SETTINGS



CLASSIFICATION REPRODUCIBLE ENABLE GPU8

SCORER

ACC
MSE
RMSE
RMSLE
MAE
GINI
ALL
LOOLOSS

ITERATION SCORES - INTERNAL VALIDATION



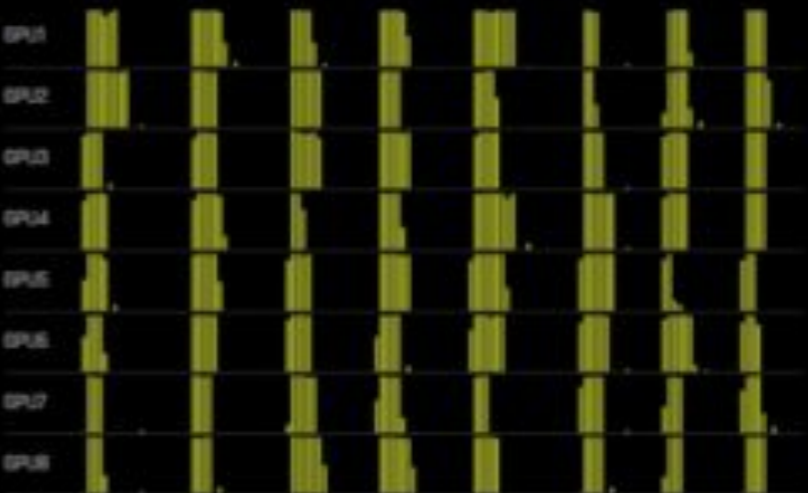
VARIABLE IMPORTANCE

89_v60	1.00
109_CV_TE_v129_v24_v66_0	0.74
107_WoE_v129_v22_v24_v30_v58_v62_v66_0	0.62
14_CV_TE_v66_0	0.30
8_WoE_v22_0	0.22
24_v10	0.18
72_NumToCatTE_v64_0	0.15
102_NumToCatTE_v30_v38_v48_v50_v74_v76_0	0.15
101_WoE_v22_v24_v56_v66_0	0.11
149_NumToCatWoE_v1_v71_v74_0	0.11
7_CV_TE_v24_0	0.10
16_CV_TE_v66_0	0.09
146_CV_CatNumInc_v66_v60_median	0.09
105_NumToCatTE_v106_v10_v24_v3_v55_v62_v66_v77_0	0.08

CPU / MEMORY



GPU USAGE



10). QLIK

- Qlik is Data mining and visualization tool. It also offers dashboards and **Supports multiple data sources and file types**
- **Features:**
 - Drag-and-drop interfaces to create flexible, interactive data visualizations
 - Instantly respond to interactions and changes.
 - Supports multiple data sources and file types
 - It allows easy security for data and content across all devices
 - It allows you to share relevant analyses, including apps and stories, using a centralized hub

[Home](#)
[Theme Editor](#)
[Gallery](#)
[About](#)
Log In

General settings

Theme name

Theme description

Inherit Sense styles?

Global font styles

Global colors

Object settings

Color picker / single color

Color by dimension

Color by measure

Preview [classic qlikview.qext](#) [theme.json](#) [theme.css](#)

Classic QlikView
QlikView It Old Skool!

Single color
Subtitle text

Footer text

Color by Dimension
Subtitle text

Color

- B
- I
- T
- M
- E
- T
- R

Footer text

Color by Measure
Subtitle text

2014												
2015												
2016												
2017												
2018												
Footer text	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec

[Download theme](#)

TECHNIQUES

➤ Data Discovery:

- **Cluster**
 - k-mean
 - DB Scan: min point & Epsilon
 - Confident interval
- **Gabor Filtering**: feature extraction: Gradient
- **Convolutional Neural Network (CNN)**
Deep Learning

➤ Prediction:

- **Data Exploration**
- **Correlation**
- **Similarity** e.g. Jacquard, Cosine Similarity

➤ Microsoft Excel

➤ MiniTab

➤ Power BI

BIBLIOGRAPHY

- [1] <http://opensourceforu.com/2017/03/top-10-open-source-data-mining-tools/>
- [2] <https://thenewstack.io/six-of-the-best-open-source-data-mining-tools/>
- [3] <http://blog.galvanize.com/four-data-mining-techniques-for-businesses-that-everyone-should-know/>
- [4] <http://www.rdatamining.com/resources/tools>
- [5] <https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques/>
- [6] <https://www.predictiveanalyticstoday.com/top-free-data-mining-software/>
- [7] <https://www.kdnuggets.com/software/index.html>
- [8] <https://www.guru99.com/best-data-mining-tools.html>