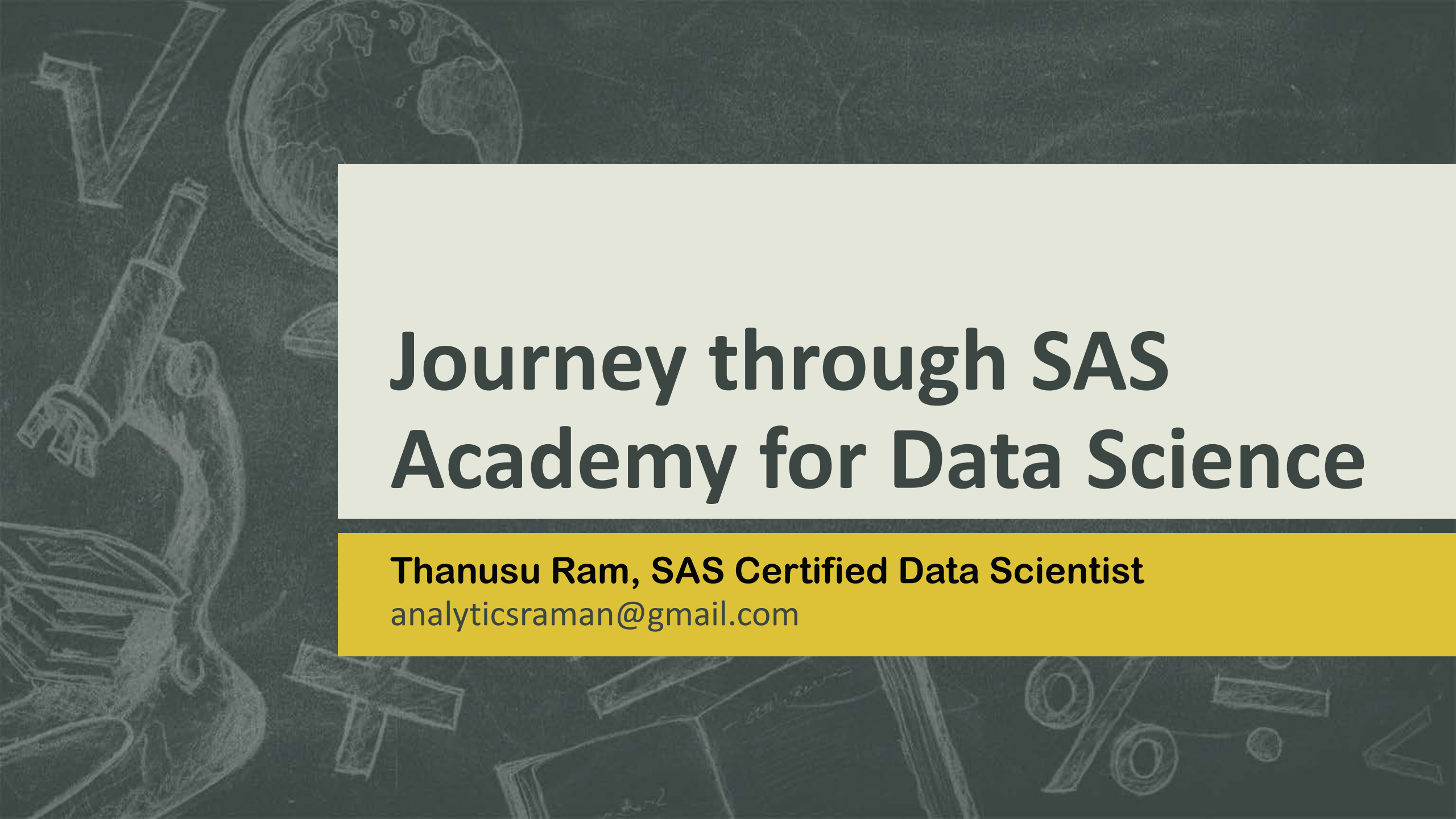


Journey through SAS Academy for Data Science

Thanusu Ram, SAS Certified Data Scientist
analyticsraman@gmail.com

Why SAS Academy for Data Science?

- Customers in 146 countries
- Installed at > 80,000 organizations, including 91 of the top 100 companies on the Fortune Global 500
- Real life case studies and projects
- Continuous progress monitoring and certifications
- Five international Certifications
- Depth of analytical content
- A blended approach that covers several programming languages

Higher degrees for Data Science? Is there an alternative?

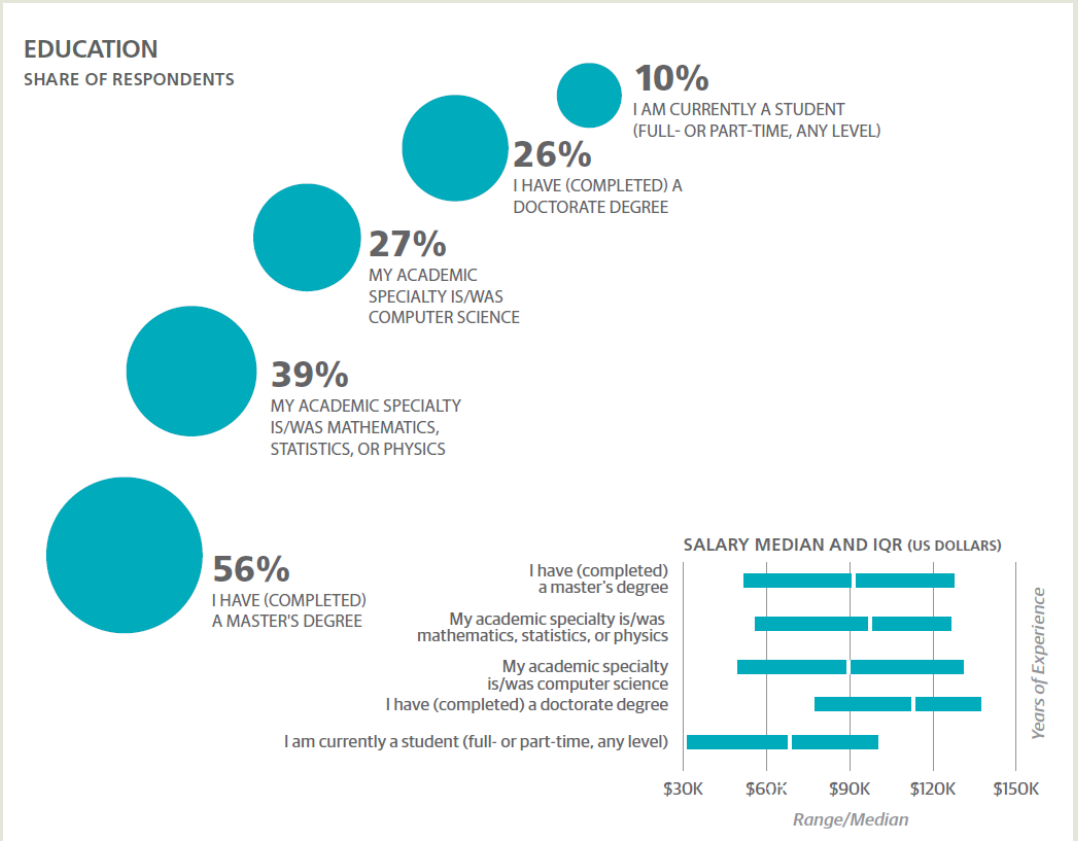
2017 DATA SCIENCE SALARY SURVEY

Education

More than 75% of respondents have a graduate degree, 56% have a master's, and 26% have a doctorate.

There is definitely an increase in salary as your degree increases. Students have a median salary of $68,000, whereas computer science majors with a degree have a salary of $89,000; those with a master's earn $91,000, and doctorates receive $113,000. You should keep in mind that just because you have a higher degree, that doesn't automatically mean that you can expect a higher wage. Having a deeper knowledge on one specific, niche topic might be in high demand, or the types of companies needing those skills might pay better, or the tasks might just be more complex and require more experience and expertise, and therefore are rewarded with higher pay.

A doctorate degree has a wage increase of around $15,000, but not entering the workforce three years earlier sets you back nearly $270,000 in lost salary, plus school tuition. How many more years would you need to work if you got an annual a $15,000 bonus to pay off that debt?



http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist/

Big Data Professional

Topics covered:

1. SAS Programming
2. Introductory Statistics
3. Data Management and Manipulation
4. Data Quality
5. Working with Hadoop, Hive and SAS
6. Data Visualization

Overview of Courses of Level 1: Big Data Professional

Big Data Challenges and Analysis-Driven Data

Provides an overview of the challenges associated with big data and analysis-driven data.

- Reading external data files.
- Storing and processing data.
- Combining Hadoop and SAS.
- Recognizing and overcoming big data challenges.

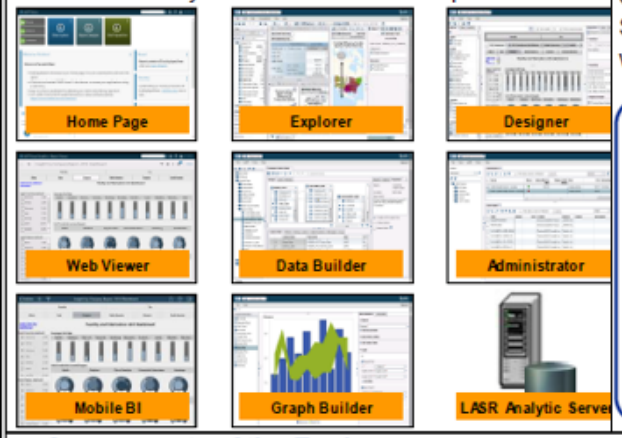
Exploring Data with SAS Visual Analytics

Using SAS Visual Analytics Explorer to explore in-memory tables from the SAS® LASR™ Analytic Server and perform advanced data analyses.

- Finding previously unknown relationships and spotting trends in your data.
- Visualizing data using charts, plots and tables.
- Using the auto charting function to visualize data in the best possible way.
- Using advanced graphs, such as network diagrams, Sankey diagrams and word clouds.
- Easily adding analytics to your graphs, and including descriptions of the analytics results.
- Navigating through your data using on-the-fly hierarchies.

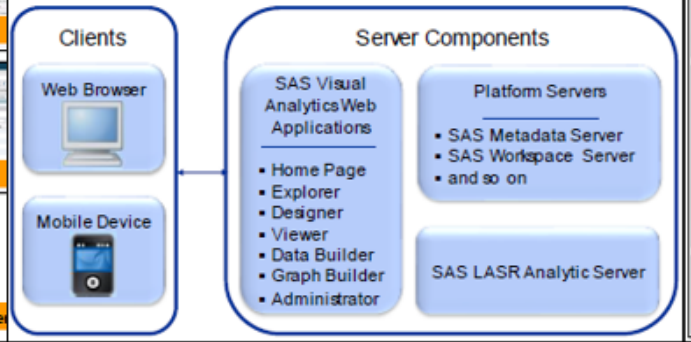
SAS Visual Analytics Architecture

SAS Visual Analytics consists of several parts.



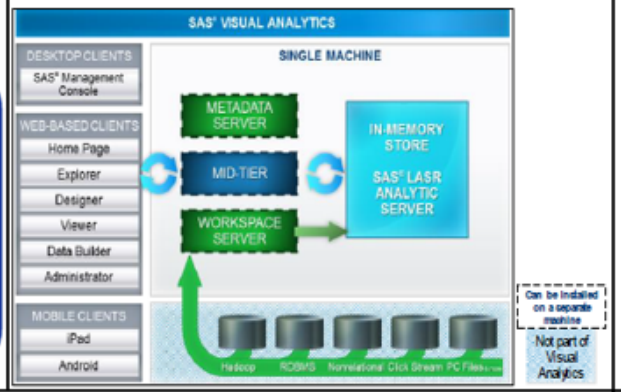
SAS Visual Analytics Server Components

SAS Visual Analytics is built on the platform for SAS Business Analytics. It is designed to work with the SAS LASR Analytic Server.



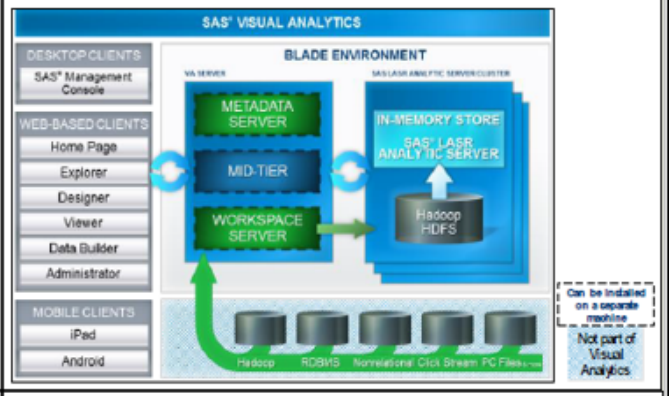
Non-Distributed Deployment

This diagram shows a typical non-distributed deployment of SAS Visual Analytics.



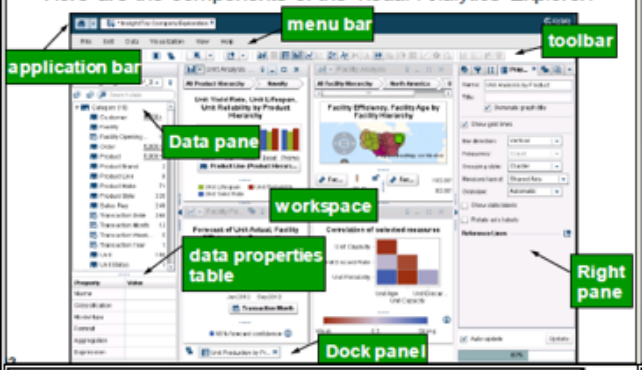
Distributed Deployment

This diagram shows a typical distributed deployment of SAS Visual Analytics.



Components of the Explorer

Here are the components of the Visual Analytics Explorer:



Visualizations and Explorations

A *visualization* displays data values using one of several visualization types. Visualization types include tables, charts, plots, geographic maps, and more. A visualization can contain filters and other display properties.

Classification Properties

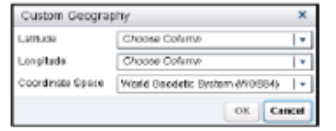
Each data item is categorized using a Classification property.

Category	Used to group and aggregate measures. Categories contain alphanumeric or datetime values. New category data items can be calculated.
Geography	Special role to identify types of geographical information for mapping.
Measure	Numeric items whose values are used in computations. Measures can be calculated or aggregated.
Hierarchy	Used to navigate through the data. Hierarchies are based on category or geography values.

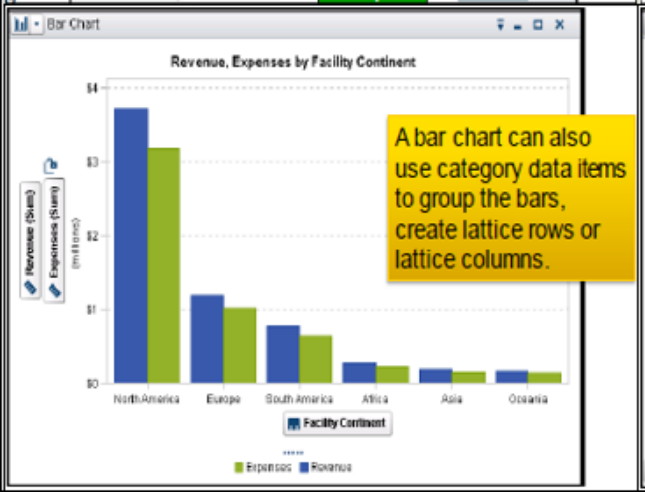
Additional data item types might be available.

Geo Map: Custom Geography

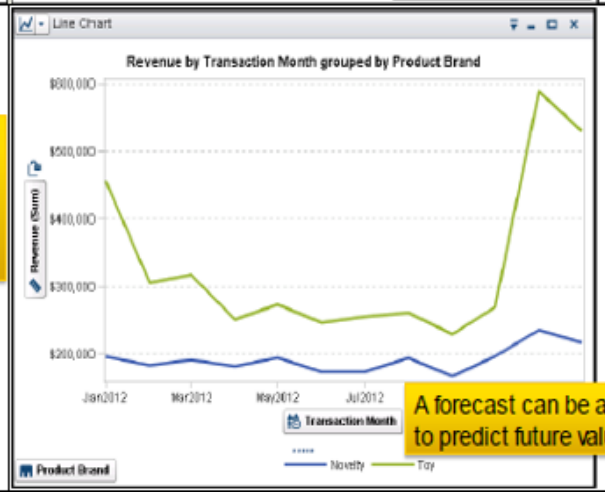
You can define a custom geography data item by providing the following information:



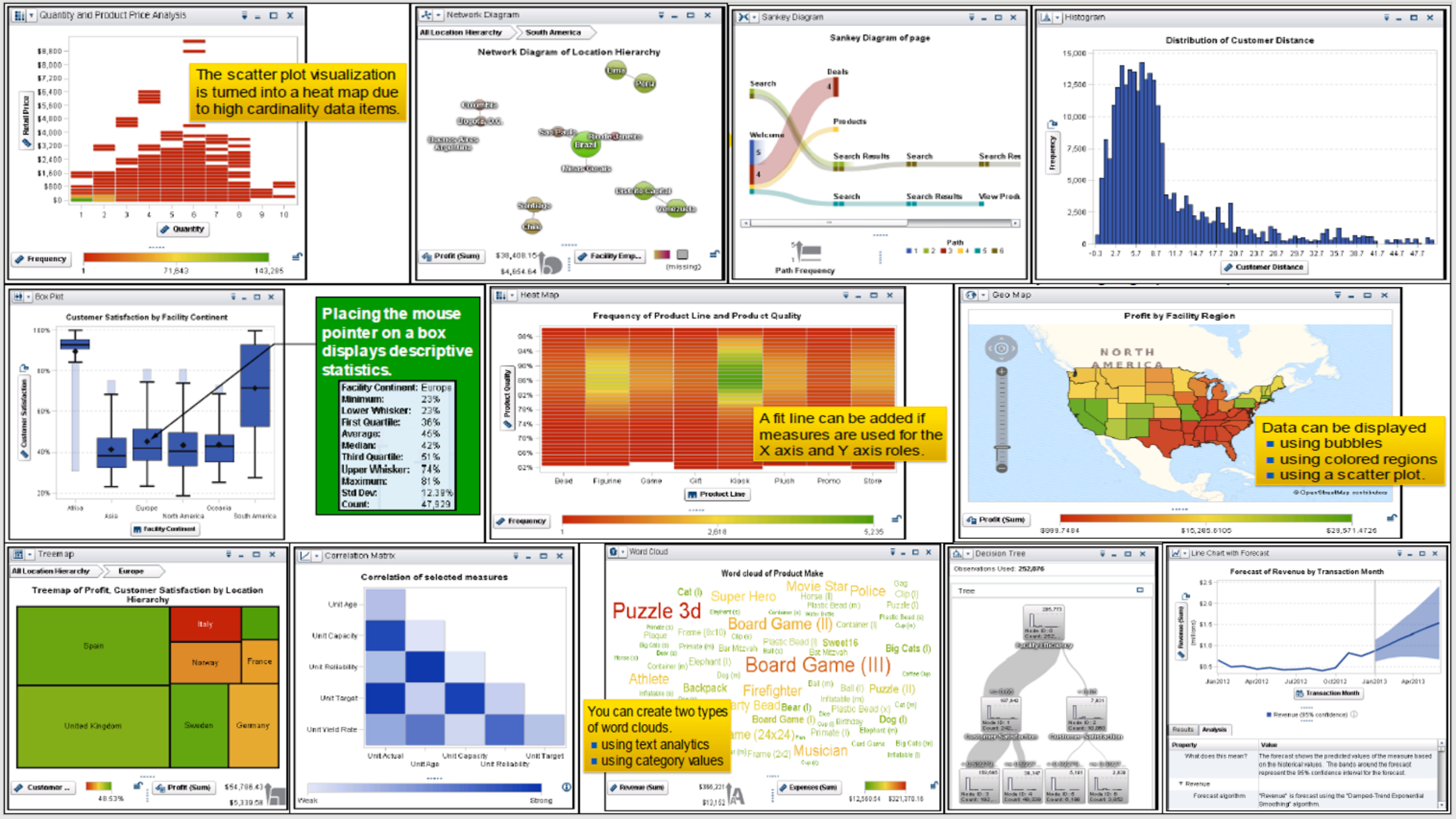
Latitude	A measure from your current data source that contains the latitude (Y) coordinate values for the geographic role that you want to define
Longitude	A measure from your current data source that contains the longitude (X) coordinate values for the geographic role that you want to define
Coordinate Space	The coordinate system that is used to project the longitude and latitude coordinates



A bar chart can also use category data items to group the bars, create lattice rows or lattice columns.



A forecast can be added to predict future values.


Quantity and Product Price Analysis
The scatter plot visualization is turned into a heat map due to high cardinality data items.
Retail Price
Quantity
Frequency
Network Diagram
All Location Hierarchy
South America
Network Diagram of Location Hierarchy
Brazil
Profit (Sum)
Facility Emp...
Sankey Diagram
Sankey Diagram of page
Search
Deals
Products
Welcome
Search Results
Path Frequency
Histogram
Distribution of Customer Distance
Frequency
Customer Distance
Box Plot
Customer Satisfaction by Facility Continent
Placing the mouse pointer on a box displays descriptive statistics.
Facility Continent: Europe
Minimum: 23%
Lower Whisker: 23%
First Quartile: 36%
Average: 45%
Median: 42%
Third Quartile: 51%
Upper Whisker: 74%
Maximum: 81%
Std Dev: 12.39%
Count: 47,929
Customer Satisfaction
Facility Continent
Heat Map
Frequency of Product Line and Product Quality
A fit line can be added if measures are used for the X axis and Y axis roles.
Product Quality
Product Line
Frequency
Geo Map
Profit by Facility Region
NORTH AMERICA
Data can be displayed
using bubbles
using colored regions
using a scatter plot.
Profit (Sum)
Treemap
All Location Hierarchy
Europe
Treemap of Profit, Customer Satisfaction by Location Hierarchy
Spain
Italy
Norway
France
United Kingdom
Sweden
Germany
Correlation Matrix
Correlation of selected measures
Unit Age
Unit Capacity
Unit Reliability
Unit Target
Unit Yield Rate
Unit Actual
Weak
Strong
Word Cloud
Word cloud of Product Make
Puzzle 3d
Board Game (II)
Board Game (III)
Athlete
Backpack
Firefighter
Musician
You can create two types of word clouds.
using text analytics
using category values
Revenue (Sum)
Expenses (Sum)
Decision Tree
Observations Used: 252,876
Tree
Line Chart with Forecast
Forecast of Revenue by Transaction Month
Revenue (Sum) (millions)
Transaction Month
Revenue (95% confidence)
Results
Analysis
Property
Value
What does this mean?
The forecast shows the predicted values of the measure based on the historical values. The bands around the forecast represent the 95% confidence interval for the forecast.
Revenue
Forecast algorithm
"Revenue" is forecast using the "Damped-Trend Exponential Smoothing" algorithm.

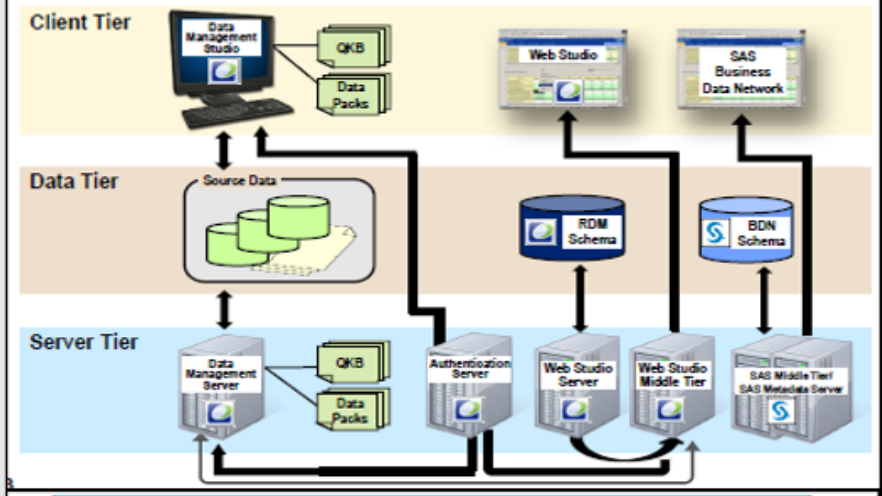
Overview of Courses of Level 1: Big Data Professional

Preparing Data for Analysis and Reporting

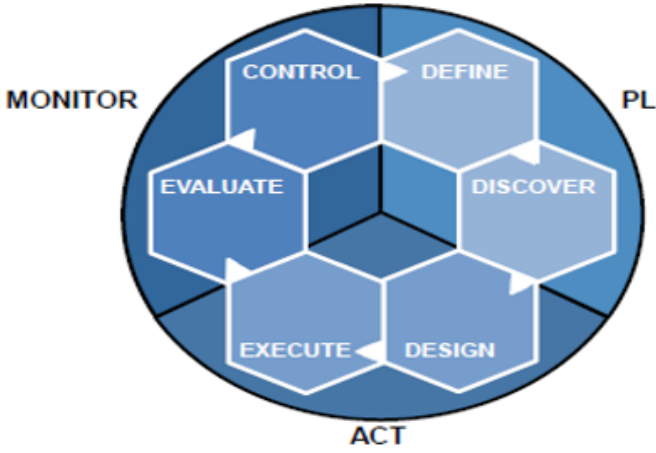
Perform data management tasks, such as improving data quality, entity resolution and data monitoring.

- Creating and reviewing data explorations.
- Creating and reviewing data profiles.
- Creating data jobs for data improvement.
- Establishing monitoring aspects for your data.
- Understanding the QKB components.
- Using the component editors.
- Understanding various definition types.
- Building a new data type (optional).

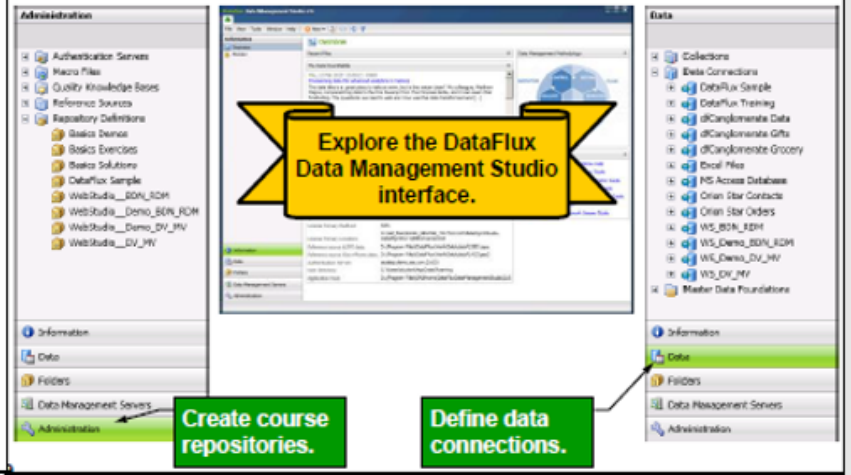
Data Management Platform Architecture



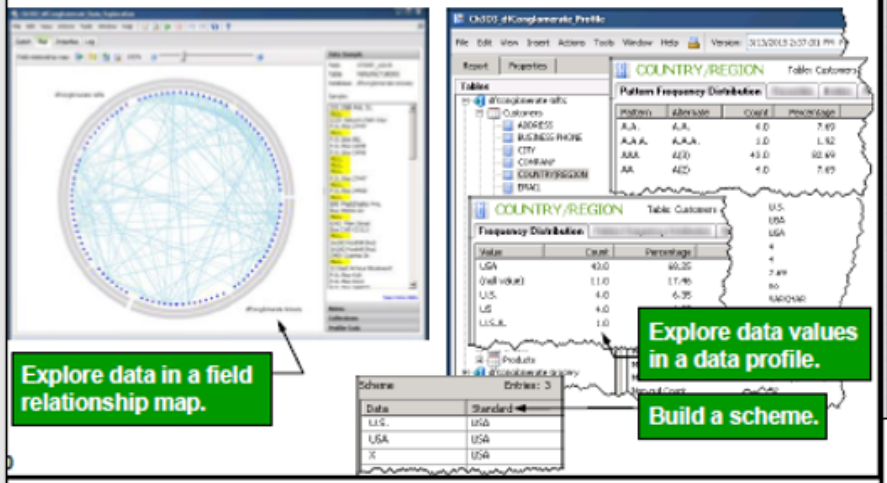
DataFlux Data Management Methodology



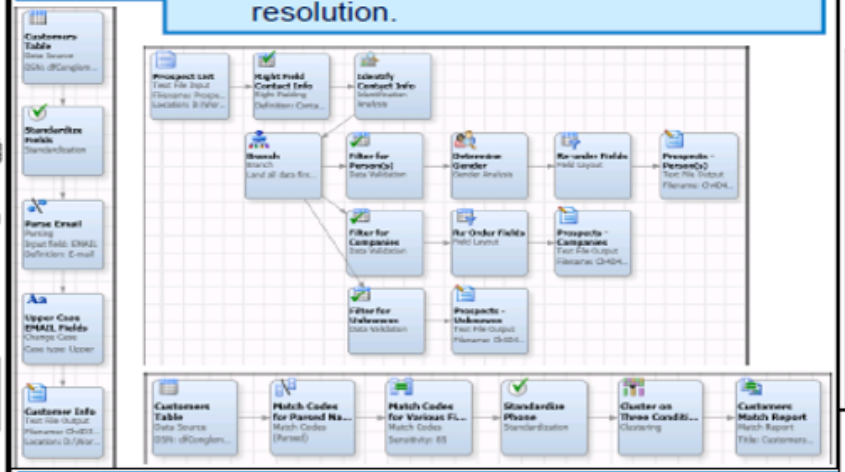
Step 1: DataFlux Data Management Studio Basics



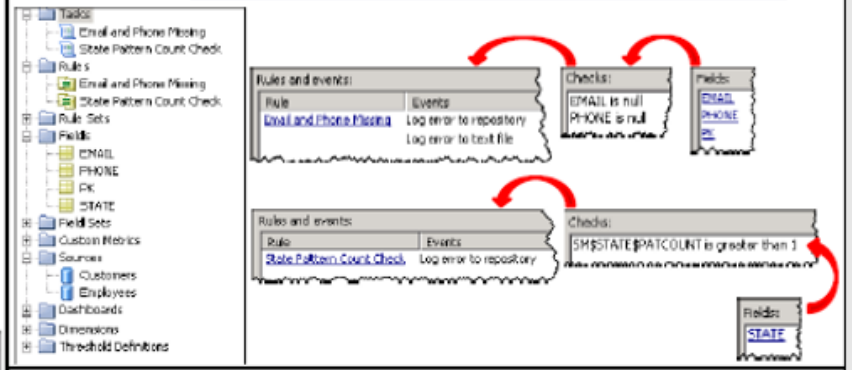
Step 2: PLAN: Connect to, explore, and profile data; design standardization schemes.



Step 3: ACT: Build data jobs for quality and entity resolution.



Step 4: MONITOR: Create and implement various types of business rules.



Step 5: Additional topics

Several additional topics are investigated.

- data joining using generated match codes
- creating and using data job references
- introducing the DataFlux Data Management Server

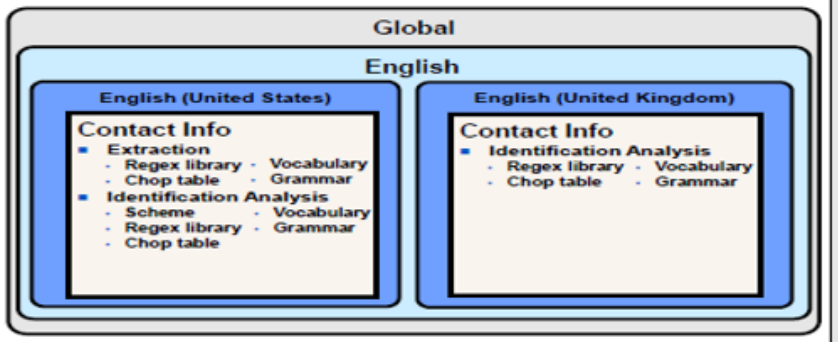
Step 6: Explore the QKB & Customization

Several additional topics are investigated.

- Why customize the QKB?
- How to explore / modify the QKB
- Investigate the various component files
- Understand the structure of the QKB (locales, data types, definitions)
- Investigate parse, standardization and match definitions

Organization of the QKB: Component Files

Each definition consists of various component files.



Overview of Courses of Level 1: Big Data Professional

Statistics 1: Introduction to ANOVA, Regression and Logistic Regression

SAS/STAT® : t-tests (one and two sample) , ANOVA (One and Two-Way), linear regression (simple and multiple), and introduction to logistic regression.

- Generating descriptive statistics and exploring data with graphs.
- Performing analysis of variance and applying multiple comparison techniques.
- Performing linear regression and assessing the assumptions.
- Using regression model selection techniques to aid in the choice of predictor variables in multiple regression.
- Using diagnostic statistics to assess statistical assumptions and identify potential outliers in multiple regression.
- Using chi-square statistics to detect associations among categorical variables.
- Fitting a multiple logistic regression model (Residuals, influential observations and collinearity).
- Scoring new data using developed models (PROC PLM & PROC GLMSELECT)

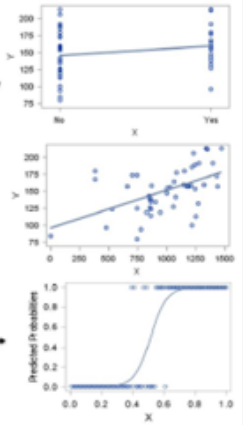
Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Logistic Regression	Logistic Regression	Logistic Regression

Overview of Models

- General Linear Models

$Y = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k + \varepsilon$

  - Analysis of Variance (ANOVA)
  - Regression



- Logistic Regression

$logit(Y) = \beta_0 + \beta_1 X_1 \ldots + \beta_k X_k$

Statistical Hypothesis Test

$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

Set null and alternative hypotheses

Set α

Choose significance level

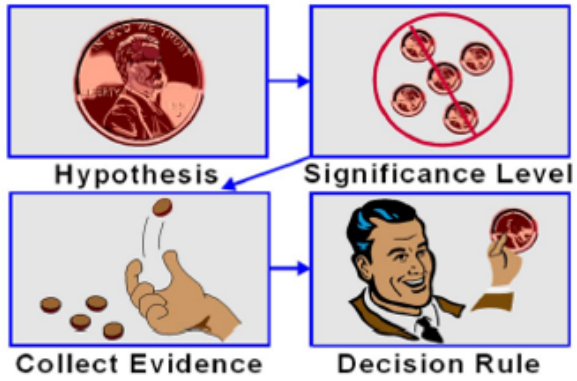
SAS

Calculate statistics

$p \geq \alpha \Rightarrow$ Do Not Reject H_0
$p < \alpha \Rightarrow$ Reject H_0

Make decision

Coin Example



What Is an Alpha Level?

You used a decision rule to make a decision, but was the decision correct?

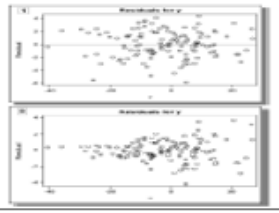
DECISION \ ACTUAL	H_0 Is True	H_0 Is False
Fail to Reject Null	Correct	Type II Error $p(\text{Type II}\mid H_1)=\beta$
Reject Null	Type I Error $p(\text{Type I}\mid H_0)=\alpha$	Correct $(1-\beta)=Power$

Chi-Square Tests

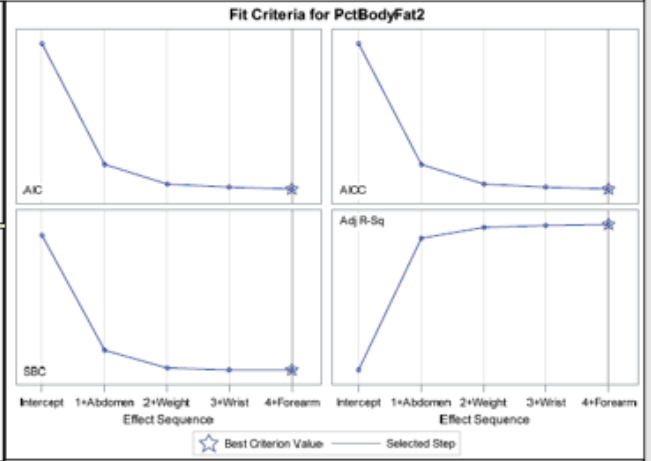
Chi-square tests and the corresponding *p*-values
- determine whether an association exists
- do not measure the strength of an association
- depend on and reflect the sample size.

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

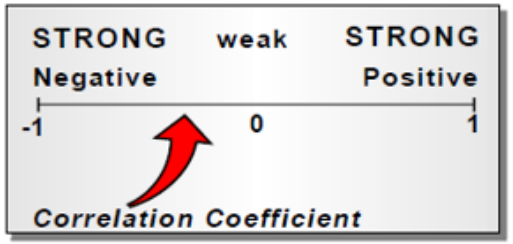
Examining Residual Plots



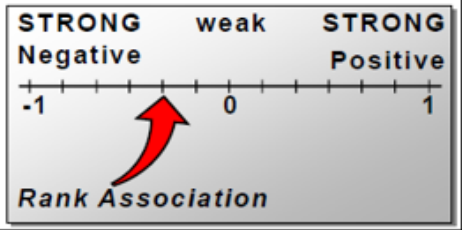
- Constant variance assumption is violated.
- Possible remedy is transforming variables to stabilize the variance.
- Procedures that model the non-constant variance can be used. (GENMOD, GLIMMIX)



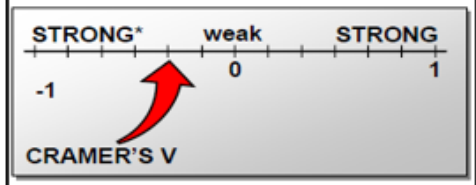
Correlation



Spearman Correlation Statistic



Measures of Association



* Cramer's V is always nonnegative for tables larger than 2*2.

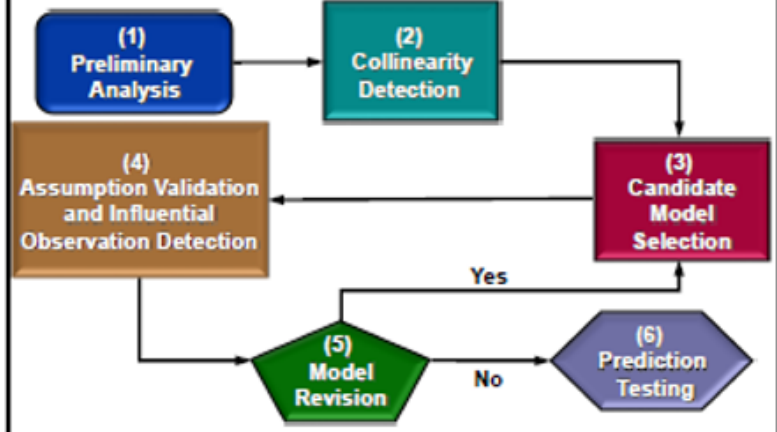
Odds Ratios

An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Example: How do the odds of irregular lot shapes being bonus eligible compare to those of regular lot shapes?

$$Odds = \frac{p_{event}}{1 - p_{event}}$$

An Effective Modeling Cycle



Overview of Courses of Level 1: Big Data Professional

Introduction to SAS and Hadoop: Essentials

Using Base SAS methods to read and write raw data with the DATA step, manage the Hadoop Distributed File System (HDFS) and execute MapReduce and Pig code from SAS via the HADOOP procedure. SAS/ACCESS® Interface to Hadoop methods that allow LIBNAME access and SQL pass-through techniques to read and write Hive or Impala table structures.

- Accessing Hadoop distributions using the LIBNAME statement and the SQL pass-through facility.
- Creating and using SQL procedure pass-through queries.
- Using options and efficiency techniques for optimizing data access performance.
- Joining data using the SQL procedure and the DATA step.
- Reading and writing Hadoop files with the FILENAME statement.
- Executing and using Hadoop commands with PROC HADOOP.
- Using Base SAS procedures with Hadoop.

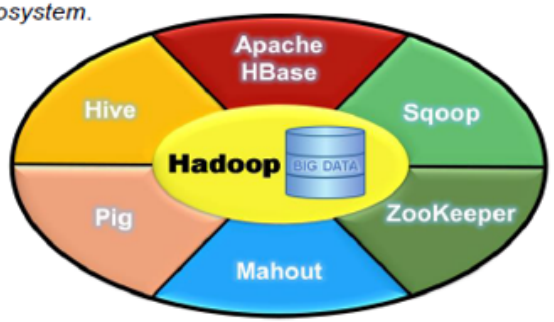
DS2 Programming Essentials with Hadoop

This course focuses on DS2, a fourth-generation SAS proprietary language for advanced data manipulation, which enables parallel processing and storage of large data with reusable methods and packages.

- Identifying the similarities and differences between the SAS DATA step and the DS2 DATA step.
- Converting a Base SAS DATA step to DS2.
- Creating DS2 variable declarations, expressions and methods for data conversion, manipulation and conditional processing.
- Creating user-defined and predefined packages to store, share and execute DS2 methods.
- Creating and executing DS2 threads for parallel processing.
- Leveraging the SAS In-Database Code Accelerator to execute DS2 code outside of a SAS session.
- Executing DS2 code in the SAS High-Performance Analytics grid using the HPDS2 procedure.

The Hadoop Ecosystem

The Apache Hadoop core technologies of HDFS, Yarn, and MapReduce, along with additional projects including Pig, Hive, and others are collectively called the *Hadoop ecosystem*.



Cloudera Hadoop Distribution Applications

Additional applications for Hadoop include:

Cloudera Manager

Deploy and manage the Hadoop Environment. This is only available with a Cloudera Distribution.

A web application that enables you to interface with Hadoop system tools. Hue is an Apache open-source application.

HUE

HDFS Viewer | Beeswax (HiveQL) | Pig Editor | Job Browser | Others

- Hortonworks, BigInsights, MapR, and other vendors can supply additional propriety technology as well. In this course, we use a Cloudera Hadoop distribution.

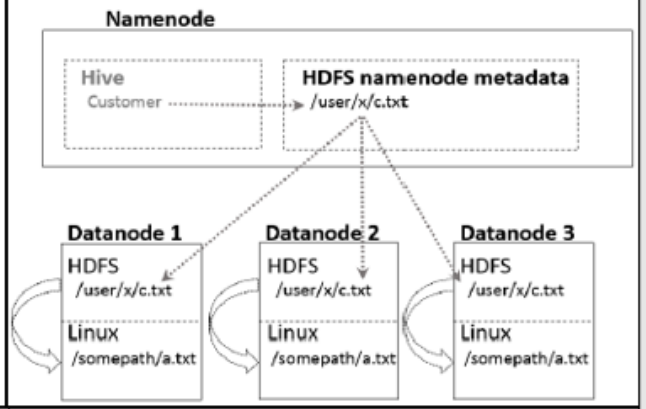
Core Hadoop Modules

Core Hadoop modules include:

HDFS (Hadoop Distributed File System)	a file system that distributes large files across the Hadoop cluster of computers
Hadoop YARN	a framework for job scheduling and cluster resource management
Hadoop MapReduce	a YARN-based system for parallel processing of large data sets

These modules automate the process of reading, writing, and processing large files in a distributed environment, freeing programmers to write programs to process the data as if they were using a single computer.

HDFS and the Linux File Systems

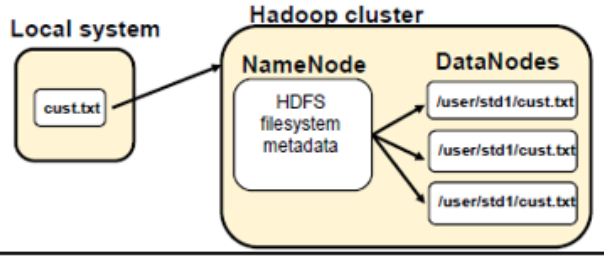


Using HDFS Commands and Files

The following HDFS command moves a local file into the HDFS cluster:

```
$ hadoop fs copyfromlocal="cust.txt"  out="/user/std1"
```

An HDFS ***NameNode*** on a root node machine distributes the file to each of the ***DataNode*** machines, and provides access to the distributed file.



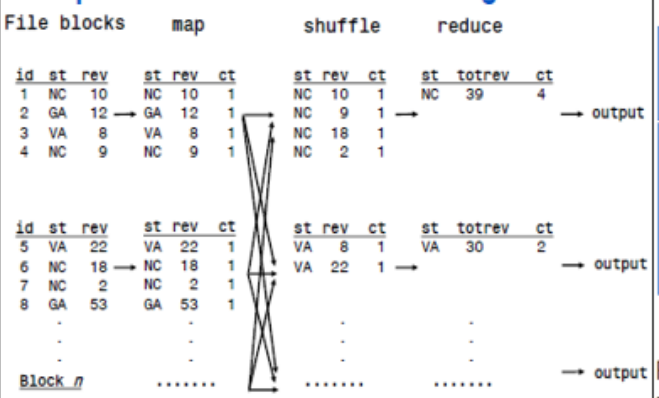
MapReduce Distributed Processing

MapReduce is a framework written in Java that is built into Hadoop. It automates the distributed processing of data files.

map	processing of individual rows (filtering, row calculations)
shuffle and sort	grouping rows for summarization
reduce	summary calculations within groups

The MapReduce framework coordinates multiple mapping, sorting, and reducing tasks that execute in parallel across the computer cluster.

MapReduce Distributed Processing



- There shuffle can produce multiple file blocks for the same value of State. The reduce step will then determine a total and a count for each block and subsequent shuffle and reduce steps could derive final statistics for each State.

Pig and Hive

Pig	A platform for data analysis that includes stepwise procedural programming that converts to MapReduce.
Hive	A data warehousing framework to query and manage large data sets stored in Hadoop. Provides a mechanism to structure the data and query the data using an SQL-like language called HiveQL. Most HiveQL queries are compiled into MapReduce programs.

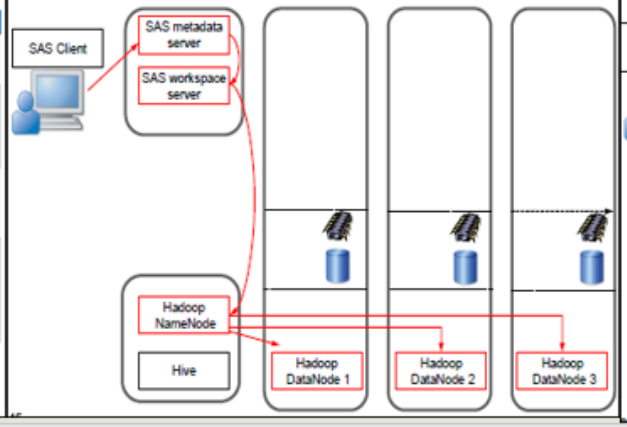
Pig and Hive provide less complex higher-level programming methods for parallel processing of Hadoop data files.

SAS Foundation Interfaces for Hadoop

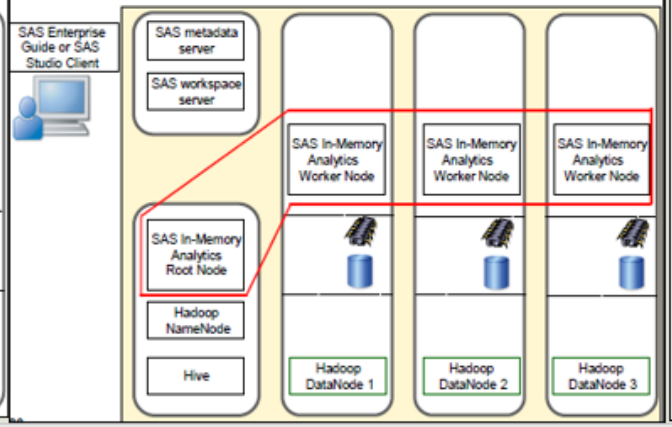
Tool	Purpose	Product
FILENAME statement	Allows the DATA step to read and write HDFS data files.	Base SAS
PROC HADOOP	Copy or move files between SAS and Hadoop. Execute Hadoop file system commands to manage files and directories. Invoke execution of existing MapReduce and Pig programs.	Base SAS
SQL Pass-Through	Submit HiveQL queries and other HiveQL statements from SAS directly to Hive for Hive processing. Query results are returned to SAS.	SAS/ACCESS Interface to Hadoop
LIBNAME Statement For Hadoop	Access Hive tables as SAS data sets using the SAS programming language. SAS/ACCESS engine translates SAS language into HiveQL and attempts to convert as much of the processing into HiveQL as possible before returning results to SAS.	SAS/ACCESS Interface to Hadoop

- If you use Cloudera Impala, you can use SAS/ACCESS Interface to Impala. It supports SQL Pass-Through and LIBNAME statement using the same methods as those for the SAS/ACCESS Interface to Hadoop shown in this course.

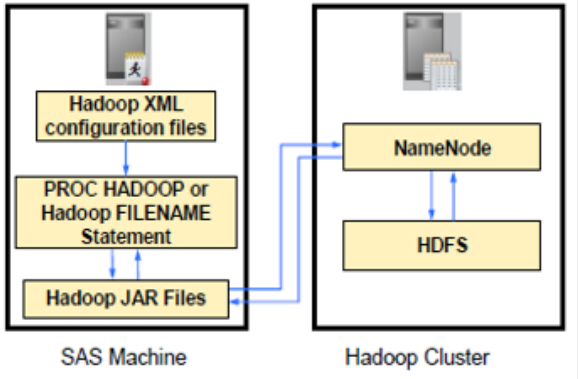
Base SAS: FILENAME for Hadoop (Review)



In-Memory Analytics



Base SAS Interface to Hadoop



Overview of Courses of Level 1: Big Data Professional

Big Data Analysis with Hive and Pig

- Moving data into the Hadoop ecosystem.
- Using Hive to design a data warehouse in Hadoop.
- Performing data analysis using HiveQL.
- Joining data sources.
- Performing ETL.
- Organizing data in Hadoop by usage.
- Performing analysis on unstructured data using Pig.
- Joining massive data sets using Pig.
- Using user-defined functions (UDFs).
- Analyzing big data in Hadoop using Hive and Pig.

Getting Started with SAS In-Memory Statistics

Accessing data on the SAS LASR Analytic Server and performing exploratory analysis and preparation.

- Starting up a SAS LASR Analytic Server.
- Loading tables into memory on the SAS LASR Analytic Server.
- Processing in-memory tables with PROC LASR and PROC IMSTAT.
- Accessing data more efficiently via intelligent partitioning.
- Deriving new temporary and permanent tables and variables.
- Creating filters and joins on in-memory data.
- Exporting ODS result tables for client-side graphic development.
- Producing descriptive statistics including counts, percentiles and means.
- Creating multidimensional summaries including cross-tabulations and contingency tables.

Business Scenario

We want to develop a prototype for a process that uses SAS to orchestrate the following scenario:

1. Move unstructured text files into the Hadoop file system.
2. Invoke MapReduce programs developed by Java programmers to:
   - read and process the text files to perform various analyses (example: word counts)
   - output results as text files in the Hadoop file system
3. Read the summarized text analysis results back into SAS for further analysis and reporting purposes.

Business Scenario Pseudocode

```
proc hadoop…;
  hdfs copyfromlocal='local file'
  out='hdfs file';
run;

proc hadoop…;
  mapreduce input='hdfs file'
            output='hdfs outfile'
            …;
run;

filename fileref 'hdfs outfile'…;
data _null_;
 infile fileref
 input …;
 …
run;
```

1. Move unstructured text files into the Hadoop file system
2. Input the hdfs file to MapReduce program and output results to hdfs file
3. Read the MapReduce output with SAS for further processing

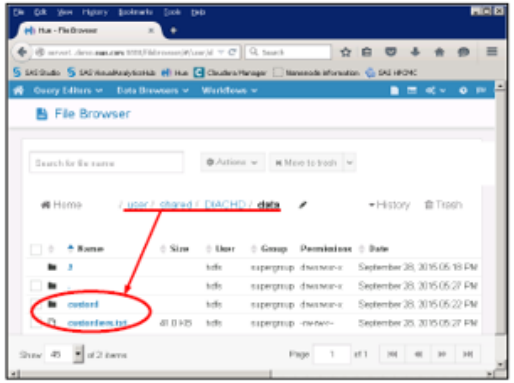
MapReduce Example

```
proc hadoop options=hadconfg username="&std" verbose;
   mapreduce
        jar=        "<linux path>/hadoop-examples-2.0.0-
                     mr1-cdh4.4.0.jar"

        input=      "<hdfs path>/moby_dick_via_sas.txt"

        map=        "org.apache.hadoop.examples.
                     WordCount$TokenizerMapper"

        reduce=     "org.apache.hadoop.examples.
                     WordCount$IntSumReducer"

        combine=    "org.apache.hadoop.examples.
                     WordCount$IntSumReducer"

        outputkey=  "org.apache.hadoop.io.Text"
        outputvalue= "org.apache.hadoop.io.IntWritable"

        output=     "<hdfs path>/mapoutput"
   ;
run;
```

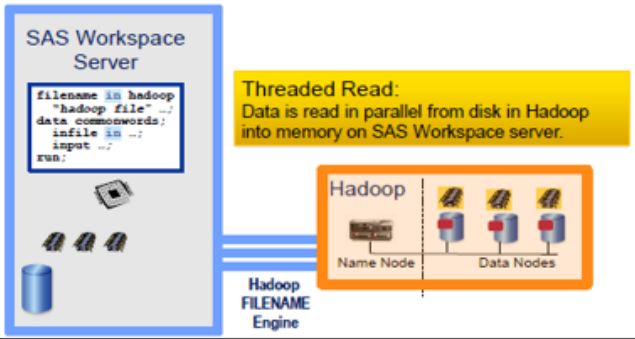
hp02d01.sas

Using Hue to Browse the Hadoop File System

The Hue application includes an HDFS interface.

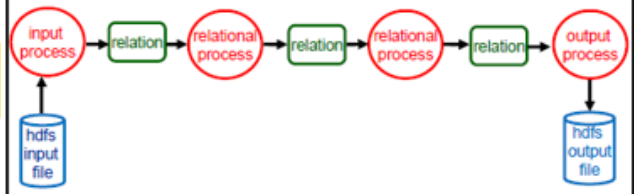


Reading a Hadoop File with a DATA Step



What Is Pig? (Self-Study)

Pig is a dataflow language that generates MapReduce to execute all of its data processes.



- not a control flow language (DO loops, IF-THEN…)
- not an object-oriented language
- not a query language

Relational Operators in Pig (Self-Study)

FOREACH	acts on each row.
FILTER	selects rows that meet a condition.
GROUP	collects rows with the same key (value).
ORDER BY	orders rows.
JOIN	joins rows in two relations by key values.
LIMIT	limits the number of rows passed to first X.
SAMPLE	samples a percentage of rows.
PARALLEL	controls number of reducer processes used for an operation.

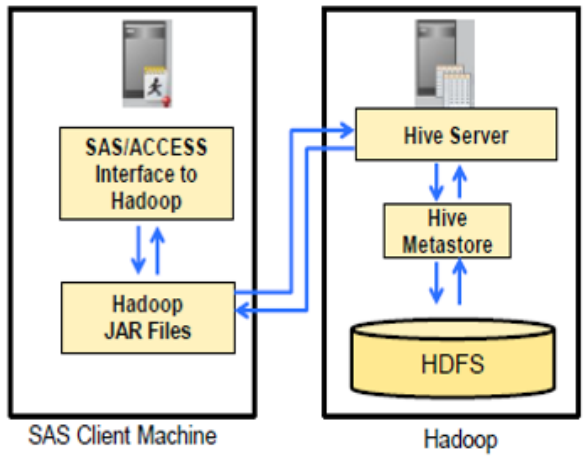
Executing Pig Code with PROC HADOOP

```
filename pigcode '/workshop/DIACHD/pigcode.txt';

proc hadoop options=hadconfg username="hdfs" verbose;
   pig code=pigcode;
run;
```

pigcode.txt:

```
A = LOAD '/user/shared/DIACHD/data/custord'
      USING PigStorage (',')
                  AS (customer_id, country, gender,
                      birth_date, product_id,
                      order_date, quantity,
                      costprice_per_unit);

B = FILTER A BY gender == 'F';

STORE B into
  '/user/student/DIACHD/data/pigoutput';
```

SAS/ACCESS Interface to Hadoop



The HiveQL Query

The query that is passed to Hive follows these rules:

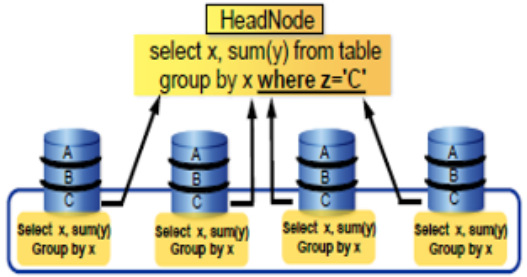
- uses SQL specific to HiveQL
- must reference Hive table and column names
- is enclosed in parentheses

```
select *
   from connection to hadoop
      (select count(*) as count
         from customer);
```

Executed by SAS

Executed by Hive

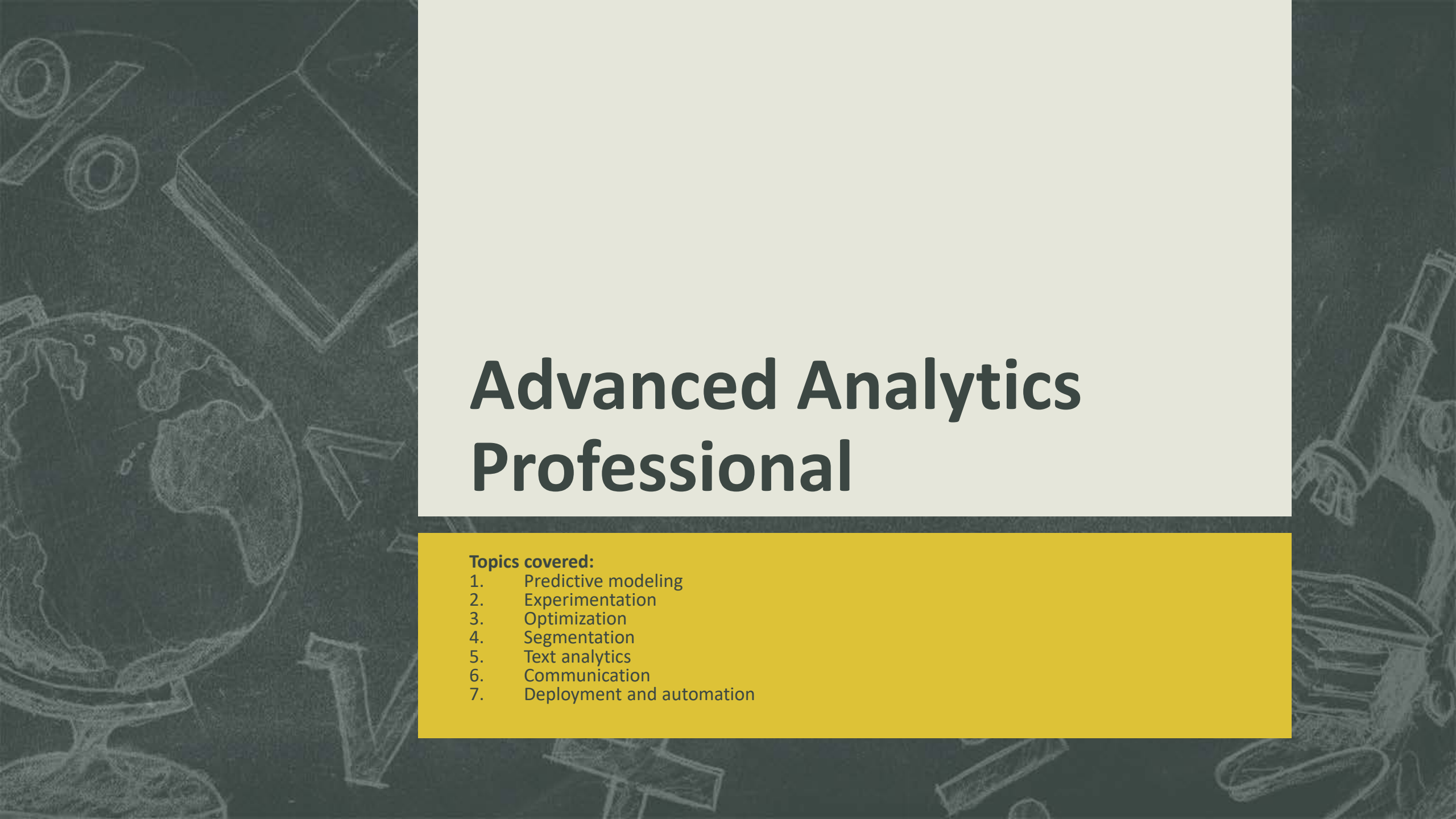
Partitions



Hadoop can find and read only the partitions that contain the selected values. This is faster and more efficient that a full table scan.

SAS In-Memory Interfaces for Hadoop

Interface	Purpose	Product
High-Performance Analytics Procedures	Perform complex analytical computations on Hadoop tables within the data nodes of the Hadoop distribution via SAS procedure language. HPDS2 allows for manipulation of data structure (column derivation).	SAS High-Performance Analytics Solutions
SAS Visual Analytics and SAS Visual Statistics	Web interfaces to generate graphical visualizations of data distributions, relationships, and analytical reports on Hadoop tables pre-loaded into memory within the data nodes of the Hadoop distribution.	SAS Visual Analytics and SAS Visual Statistics
PROC IMSTAT	A programming interface to perform complex analytical calculations on Hadoop tables pre-loaded into memory within the data nodes of the Hadoop distribution.	SAS In-Memory Statistics
DS2	A SAS proprietary language for table manipulation that translates to database language and executes in parallel in the data nodes of a distributed database.	SAS In-Database Code Accelerators



Advanced Analytics Professional

Topics covered:

1. Predictive modeling
2. Experimentation
3. Optimization
4. Segmentation
5. Text analytics
6. Communication
7. Deployment and automation

Overview of Courses of Level 2: Advanced Analytics Professional

Applied Analytics Using SAS Enterprise Miner

• Defining a SAS Enterprise Miner project, exploring data graphically and assemble analysis flow diagrams using SAS Enterprise Miner

• Modifying data for better analysis results.

• Building and understanding predictive models, including decision trees, regression and neural network models

• Comparing and explaining complex models.

• Generating and using score code.

• Applying association and sequence (pattern) discovery to transaction data.

Communicating Technical Findings to a Non-Technical Audience

• Diagnosing and assessing different styles of human behavior.

• Communicating and coping more effectively with different types of people.

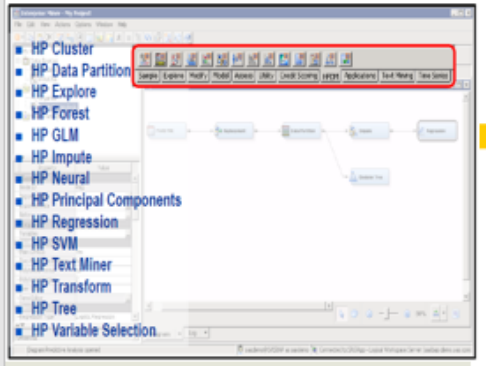
• Using your own strengths and knowledge of others to enhance communication.

• Delivering information in a concise and well-organized format.

• Creating a presentation, with the focus on communicating unfamiliar or technical information to a nontechnical audience.

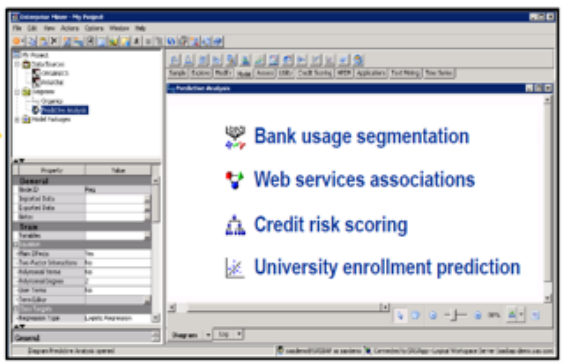
• Designing presentation materials with clarity and purpose.

Beyond SEMMA – HPDM Tab

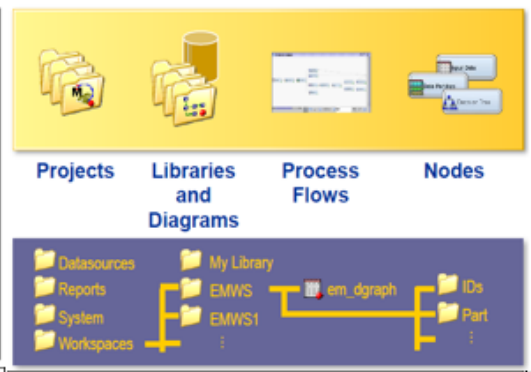


The Analytic Workflow

Applied Analytics Case Studies



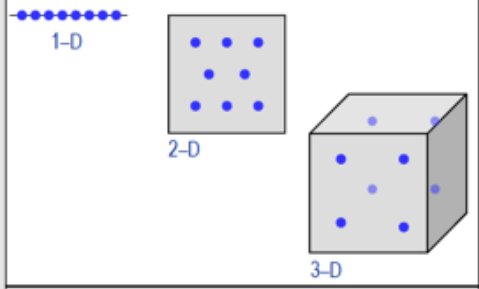
Analysis Element Organization



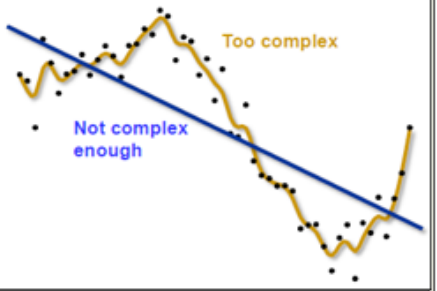
Predictive Modeling Applications

- Database marketing
- Financial risk management
- Fraud detection
- Process monitoring
- Pattern detection

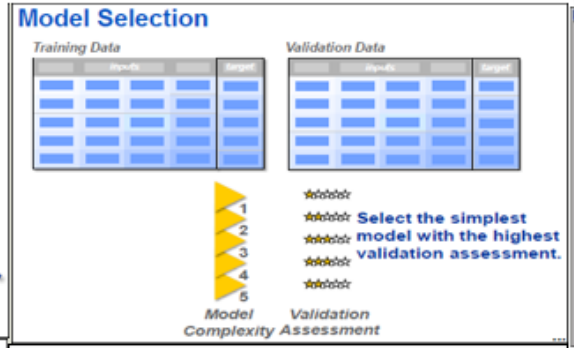
The Curse of Dimensionality

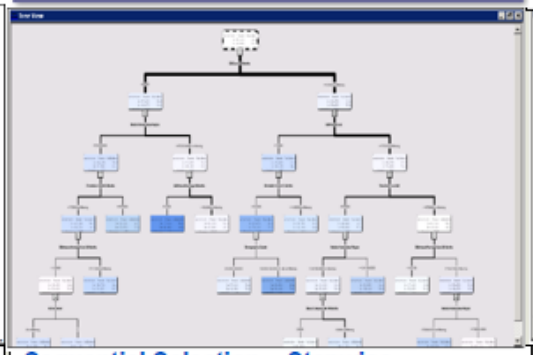


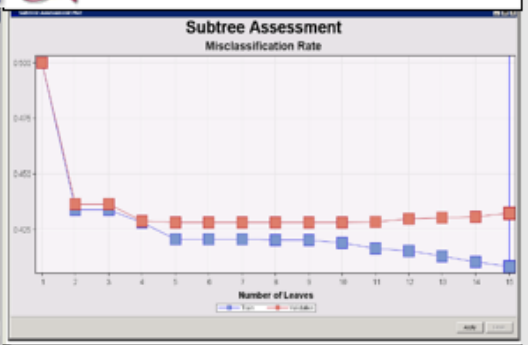
Model Complexity



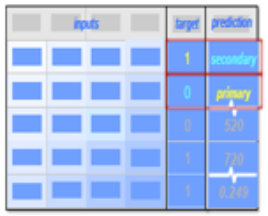
Model Selection







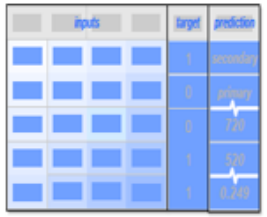
Decision Optimization: Misclassification



false negative
false positive

Minimize *misclassification*: disagreement between outcome and prediction

Complexity Optimization: Summary



decisions
accuracy / misclassification

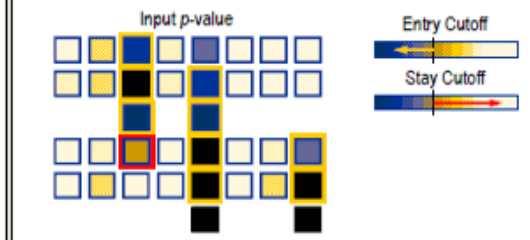
rankings
concordance / discordance

estimates
squared error

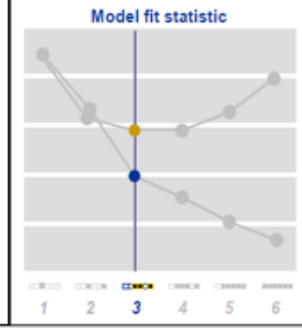
Regressions: Beyond the Prediction Formula

- Manage missing values.
- Interpret the model.
- Handle extreme or unusual values.
- Use nonnumeric inputs.
- Account for nonlinearities.

Sequential Selection – Stepwise



Select Model with Optimal Validation Fit



Evaluate each sequence step.

Choose simplest optimal model.

Odds Ratios and Doubling Amounts

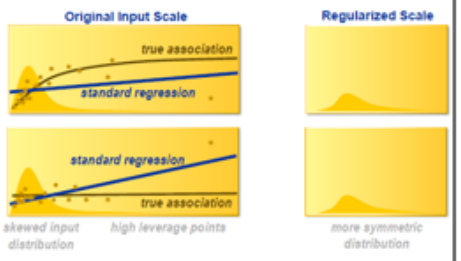
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$ logit scores

Δx_i	consequence
1	$\Rightarrow$ odds $\times \exp(w_i)$
$\frac{0.69}{w_i}$	$\Rightarrow$ odds $\times 2$

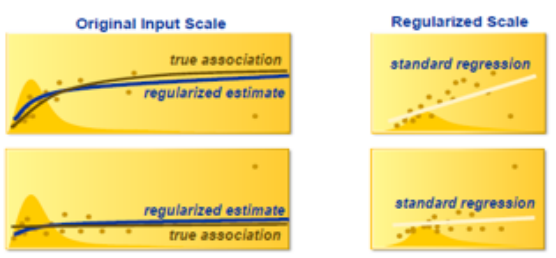
Odds ratio: Amount odds change with a unit change in input.

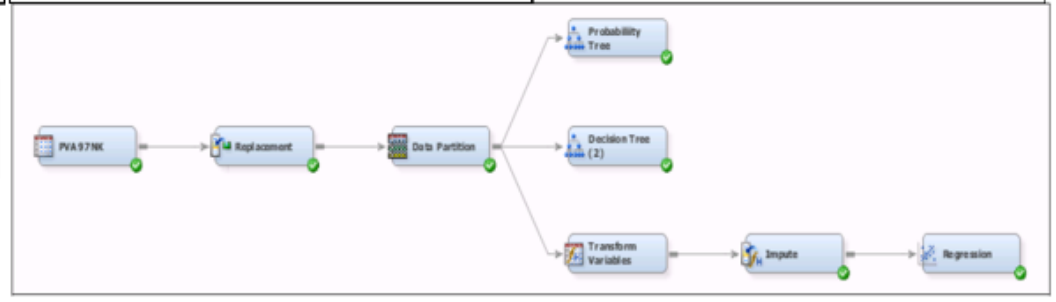
Doubling amount: How much does an input need to change to double the odds?

Extreme Distributions and Regressions



Regularizing Input Transformations





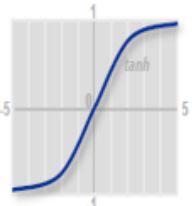
Model Essentials – Neural Networks

- Predict new cases. → Prediction formula
- Select useful inputs. → None
- Optimize complexity. → Stopped training

Neural Network Prediction Formula

$$\hat{y} = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$

prediction estimate; hidden unit; bias estimate; weight estimate



$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$
$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$
$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

activation function

Neural Network Diagram

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$

input layer; hidden layer; target layer

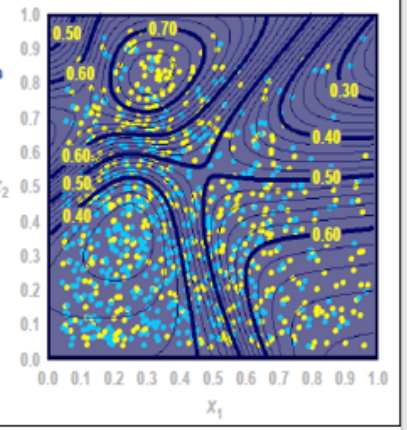
$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$
$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$
$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

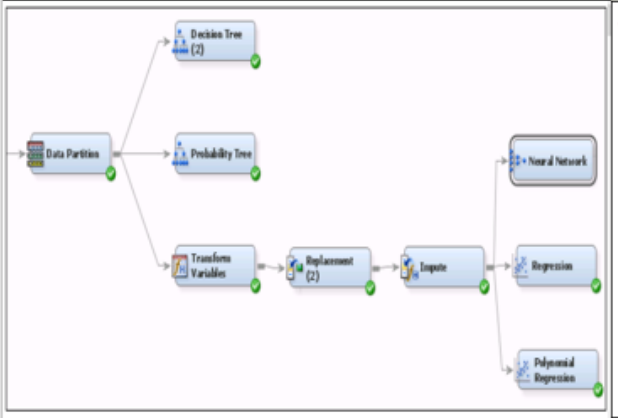
Prediction Illustration – Neural Networks

logit equation

$$\text{logit}(\hat{p}) = -0.5 + -2.6\, H_1 + -1.9\, H_2 + 0.63\, H_3$$
$$H_1 = \tanh(-1.8 + 0.25\, x_1 + -1.8\, x_2)$$
$$H_2 = \tanh(2.7 + 2.7\, x_1 + -5.3\, x_2)$$
$$H_3 = \tanh(-5.0 + 8.1\, x_1 + 4.3\, x_2)$$
$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

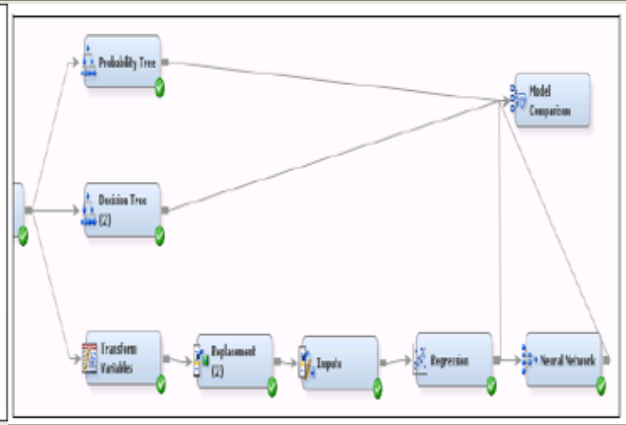
Probability estimates are obtained by solving the logit equation for $\hat{p}$ for each (x_1, x_2).





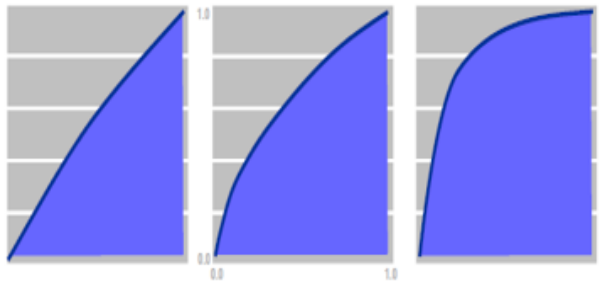
Summary Statistics Summary

Prediction Type	Statistic
Decisions	Accuracy/Misclassification Profit/Loss Inverse prior threshold
Rankings	ROC Index (concordance) Gini coefficient
Estimates	Average squared error SBC/Likelihood



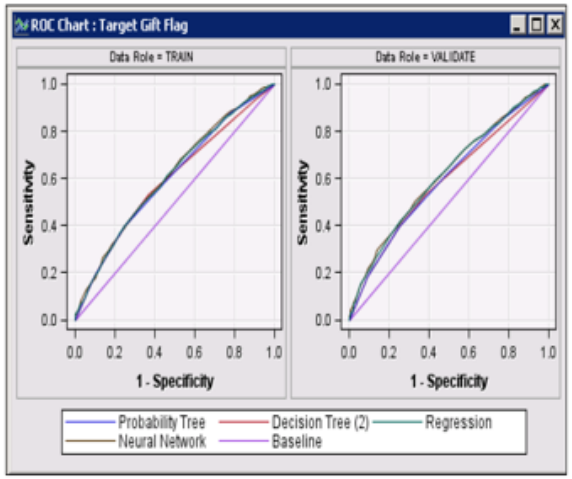
Prediction Type	Validation Fit Statistic	Direction
Decisions	Misclassification	smallest
	Average Profit/Loss	largest/smallest
	Kolmogorov-Smirnov Statistic	largest
Rankings	ROC Index (concordance)	largest
	Gini Coefficient	largest
Estimates	Average Squared Error	smallest
	Schwarz's Bayesian Criterion	smallest
	Log-Likelihood	largest

Statistical Graphics – ROC Index

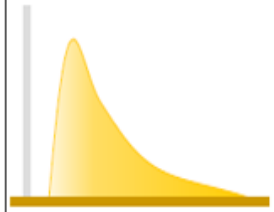


weak model
ROC Index < 0.6

strong model
ROC Index > 0.7



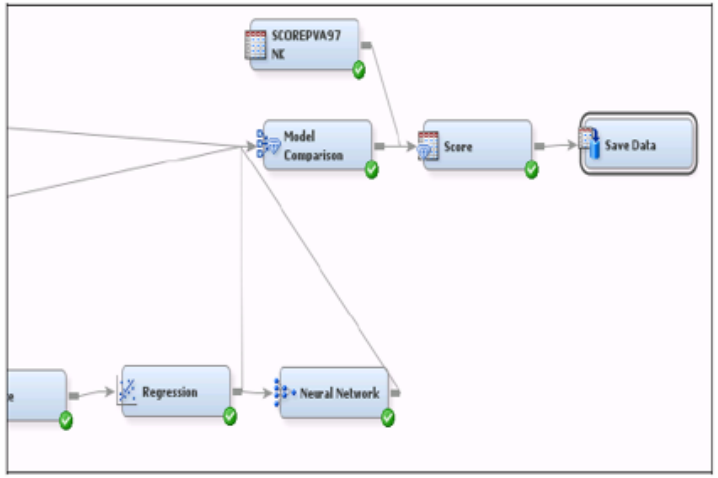
Profit Matrices



	solicit	ignore
primary outcome	15.14	0
secondary outcome	-0.68	0

0

profit distribution for solicit decision



Overview of Courses of Level 2: Advanced Analytics Professional

Neural Network Modeling

• Constructing multilayer perceptron and radial basis function neural networks.

• Constructing custom neural networks using the NEURAL procedure.

• Choosing an appropriate network architecture and determining the relevant training method.

• Avoiding overfitting neural networks.

• Performing autoregressive time series analysis using neural networks.

• Interpreting neural network models

Predictive Modeling Using Logistic Regression

• Using logistic regression to model an individual's behavior as a function of known inputs.

• Selecting variables and interactions.

• Creating effect plots and odds ratio plots using ODS Statistical Graphics.

• Handling missing data values.

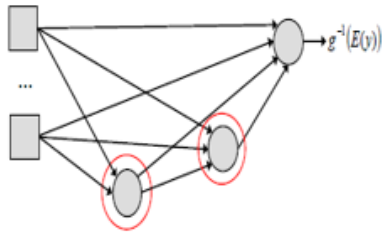
• Tackling multicollinearity in your predictors.

• Assessing model performance and comparing models.

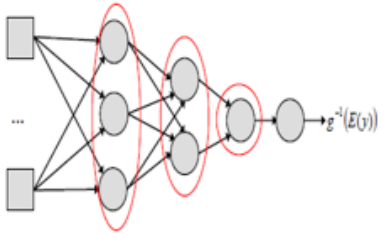
• Recoding categorical variables based on the smooth weight of evidence.

• Using efficiency techniques for massive data sets.

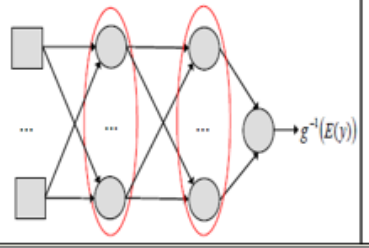
Cascade Architecture



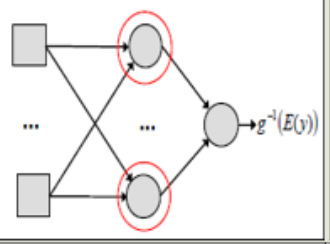
Funnel Layers Architecture



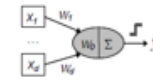
Block Layers Architecture



Single-Layer Architecture



The McCulloch-Pitts Neuron



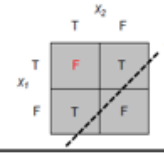
- A McCulloch-Pitts neuron is defined by the equation:

$$\hat{y} = f\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)$$

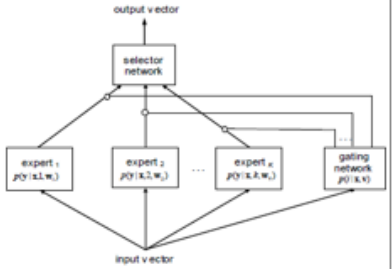
- The step function, $f(.)$, turns each neuron into a linear classifier/discriminator.

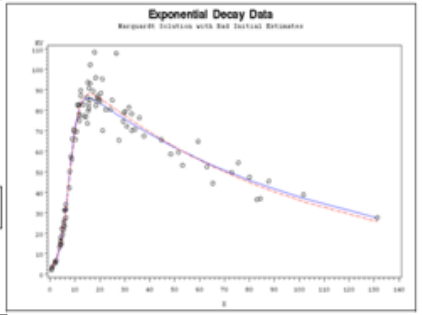
Limitations of Rosenblatt's Perceptron

- Minsky and Papert found that Rosenblatt's perceptron could solve only linearly separable problems.
- The exclusive-or truth table (see below) is an example of a problem that is not linearly separable.



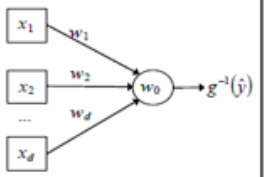
Mixture of Experts Networks



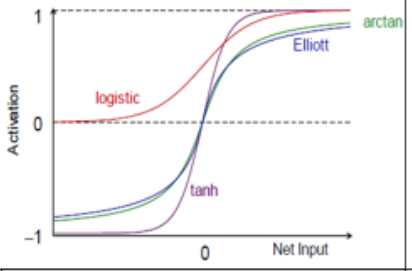


Linear Perceptron

$$g^{-1}(\hat{y}) = w_0 + \sum_{i=1}^{d} w_i x_i$$



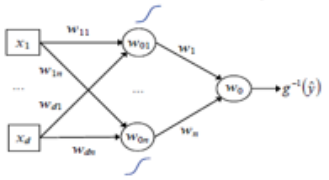
Activation Functions



Multilayer Perceptron

$$g^{-1}(\hat{y}) = w_0 + \sum_{i=1}^{h} w_i g_i\left(w_{0i} + \sum_{j=1}^{d} w_{ij} x_j\right)$$

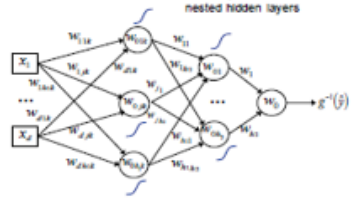
hidden layer



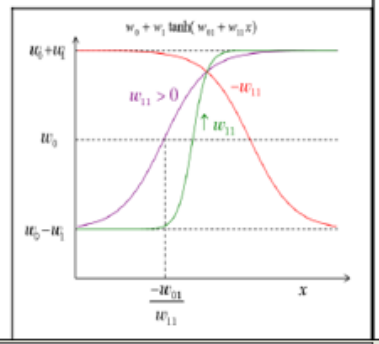
MLP with Two Hidden Layers

$$g^{-1}(\hat{y}) = w_0 + \sum_{k=1}^{h_2} w_k g_k\left(w_{0k} + \sum_{j=1}^{h_1} w_{jk}\, g_j\left(w_{0j} + \sum_{i=1}^{d} w_{ij} x_i\right)\right)$$

nested hidden layers

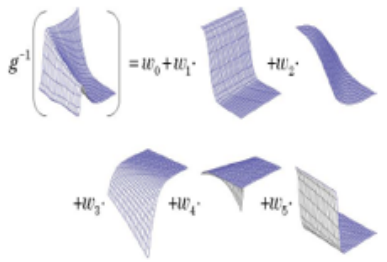


Shaping the Sigmoid



Sigmoidal Basis Functions

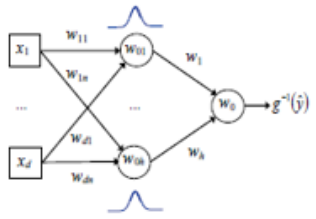
By combining (adding) the weighted sigmoids from the hidden layer neurons, any shape can be modeled.



Ordinary Radial Basis Function

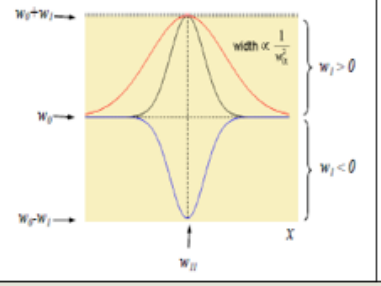
$$g^{-1}(\hat{y}) = w_0 + \sum_{i=1}^{h} w_i \exp\left[-w_{0i}\left(\sum_j (w_{ij} - x_j)^2\right)\right]$$

hidden unit



Shaping the Gaussian

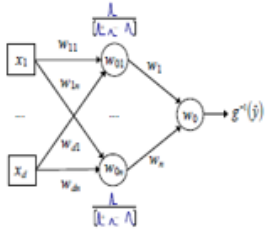
$$w_0 + w_1 \exp\left(-w_{01}^2 (x - w_{11})^2\right)$$



Normalized Radial Basis Function

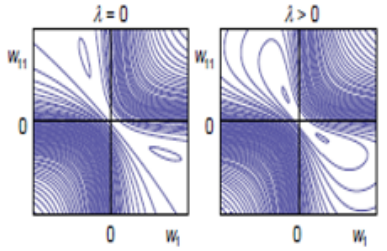
$$g^{-1}(\hat{y}) = w_0 + \sum_{i=1}^{h} w_i \operatorname{softmax}\left\{f \cdot \ln(a_i) - w_{0i}^2\left(\sum_j (w_{ij} - x_j)^2\right)\right\}$$

hidden unit

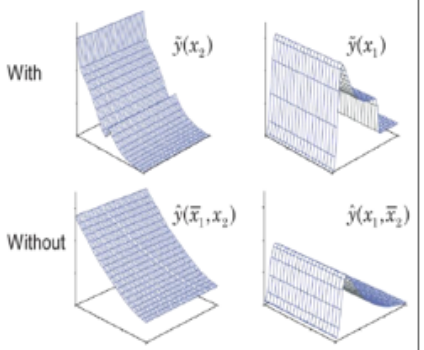


Weight Decay

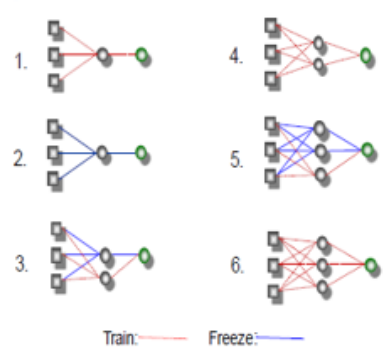
Objective Function = Error Function + $\lambda \|\mathbf{w}\|^2$



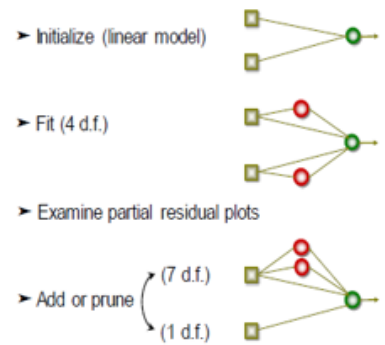
Sensitivity-Based Pruning



Sequential Network Construction

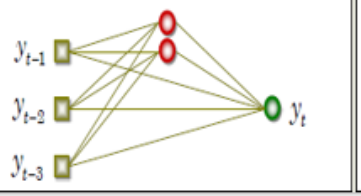


GANN Methodology

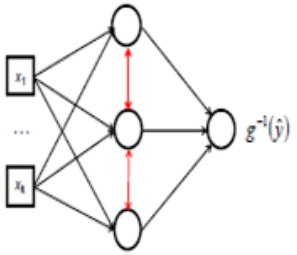


Time Delay Neural Network

$$g_0^{-1}(E(y_t)) = w_0 + w_1 y_{t-1} + w_2 y_{t-2} + w_3 y_{t-3} + w_4 \tanh\left(w_{01} + w_{11} y_{t-1} + w_{21} y_{t-2} + w_{31} y_{t-3}\right) + w_5 \tanh\left(w_{02} + w_{12} y_{t-1} + w_{22} y_{t-2} + w_{32} y_{t-3}\right)$$



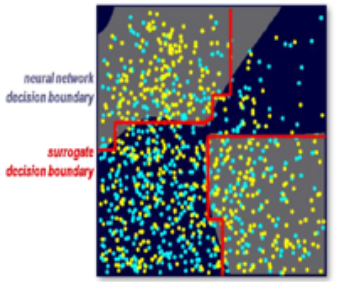
Counterpropagation



Input Layer Hidden Layer (Kohonen SOM) Output Layer

Surrogate Models

A surrogate model approximates an inscrutable model's predictions/decisions in order to facilitate interpretation.



Overview of Courses of Level 2: Advanced Analytics Professional

Data Mining Techniques: Predictive Analytics on Big Data

It presents basic and advanced modeling strategies, such as group-by processing for linear models, random forests, generalized linear models and mixture distribution models.

- Using applications (as SAS Enterprise Miner, SAS Visual Statistics and SAS In-Memory Statistics) designed for big data analyses (assaying and modeling).
- Exploring data efficiently.
- Reducing data dimensionality.
- Building predictive models using decision trees, regressions, generalized linear models, random forests and support vector machines.
- Building models that handle multiple targets.
- Assessing model performance.
- Implementing models and scoring new predictions.

Using SAS to Put Open Source Models into Production

Topics are presented in the context of data mining, which includes data exploration, model prototyping, and supervised and unsupervised learning techniques.

- Calling R packages in SAS.
- Leveraging Python scripts in SAS.
- Integrating open source data exploration techniques in SAS.
- Integrating open source models in SAS Enterprise Miner.
- Creating production (score) code for R models.

Overview of Courses of Level 2: Advanced Analytics Professional

Text Analytics Using SAS Text Miner

• Converting documents stored in standard formats (Microsoft Word, Adobe PDF, etc.) into general-purpose HTML or TXT formats.

• Reading documents from a variety of sources (web pages, flat files, data elements in a relational database, spreadsheet cells, etc.) into SAS tables.

• Processing textual data for text mining (e.g., correcting misspellings or recoding acronyms and abbreviations).

• Converting unstructured text-based character data into structured numeric data.

• Exploring words and phrases in a document collection.

• Querying document collections using keywords (i.e., identifying documents that include specific words or phrases).

• Identifying topics or concepts that appear in a document collection.

• Creating user-influenced topic tables from scratch or by modifying machine-generated topics, or creating concepts using domain knowledge.

• Using derived topic tables or pre-existing user-influenced topic tables (or both) to enhance information retrieval and document classification.

• Clustering documents into homogeneous subgroups.

• Classifying documents into predefined categories.

• Integrate text data with structured data to enrich predictive modeling endeavors.

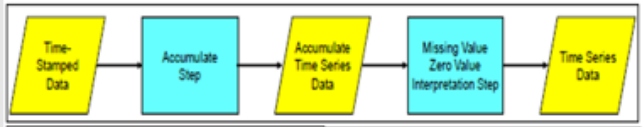
Time Series Modeling Essentials

• Creating time series data.

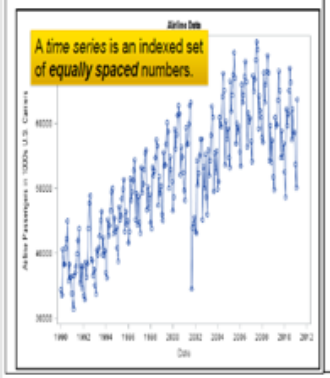
• Accommodating trend, as well as seasonal and event-related variation, in time series models.

• Diagnosing, fitting and interpreting exponential smoothing, ARIMAX and UCM models - Analyzing univariate time series

• Identifying relative strengths and weaknesses of the three model types.

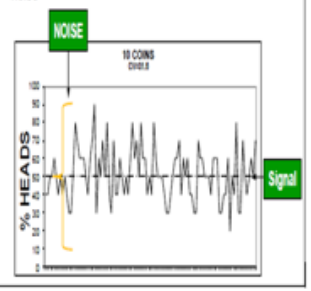


A Time Series

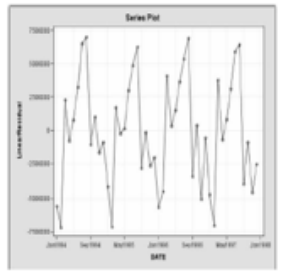


Variation in Time Series Data: Two Main Parts

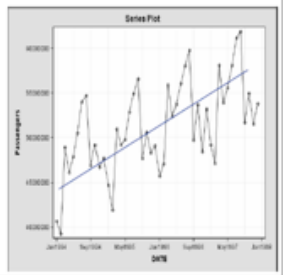
- signal
- noise



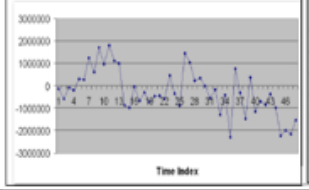
Signal Components: Seasonality



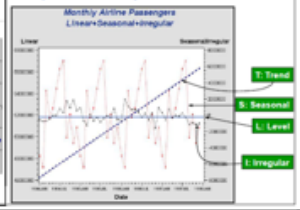
Signal Components: Trend



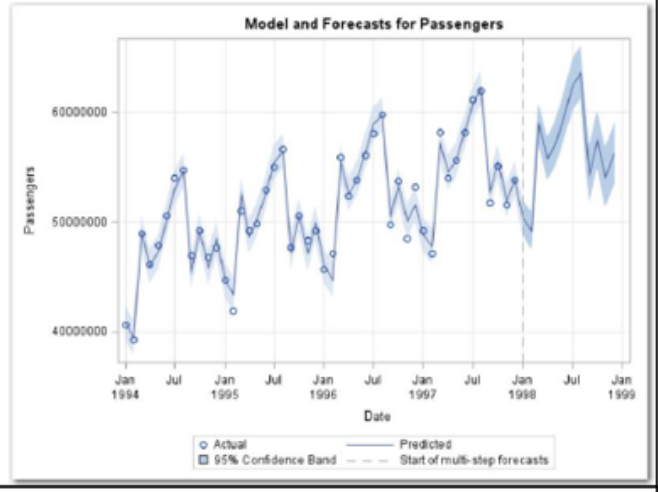
Signal Components: Irregular



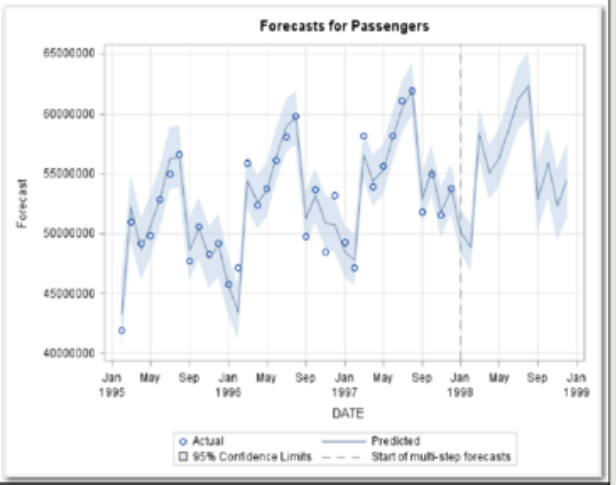
Component Decomposition



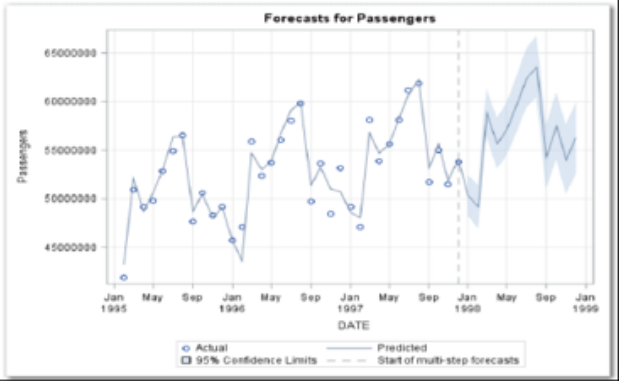
Exponential Smoothing Model Forecasts



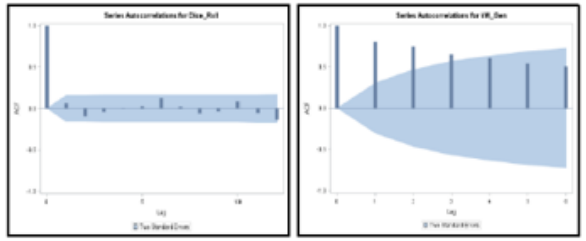
Box-Jenkins ARIMAX Forecasts



UCM Forecasts

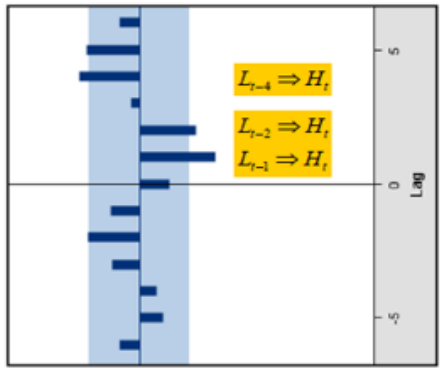


Autocorrelation Plots



- The autocorrelation plot enables you to see the autocorrelation at multiple lags.
- The blue range indicates 95% confidence intervals for each lag.

The Cross-Correlation Function (CCF)



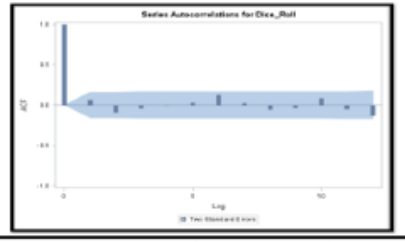
L
Jan
Feb
Mar
Apr
May

H
Jan
Feb
Mar
Apr
May

White Noise

- a series that varies randomly around its mean
- no systematic variation
- comprised of only random variation

Forecasting a white noise process reverts to the mean.



Mean 7.02

Yt	Count
2	5
3	3
4	15
5	21
6	21
7	21
8	21
9	18
10	11
11	12
12	2

The Ljung-Box Chi-Square Test for White Noise

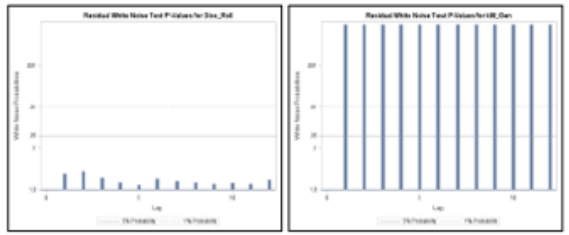
- A *white noise* time series is a Gaussian (normal, bell-shaped) time series with mean zero and positive fixed variance in which all observations are independent of each other.
- The null hypothesis is that the series is white noise, and the alternative hypothesis is that one or more autocorrelations up to lag *m* are not zero.

  H_0: The series is white noise.

  H_1: The series is *not* white noise.

- The Ljung-Box test can be applied to the original series or to the residuals after fitting a model.

The Ljung-Box Chi-Square Test for White Noise



"White means white."

ARMA Models: Initial Process Flow

Begin with a stationary series.

Examine the ACF and perform a formal test of white noise

Is the series white noise?

Yes

Tough luck – There is no systematic variation in the series. Your best guess is the mean of the series.

No

Great – The series is *not* white noise. The systematic variation in the data can be used. ARMA models are appropriate.

Forecasting Solar Power Production

Overview of Courses of Level 2: Advanced Analytics Professional

Experimentation in Data Science

• Defining common terminology in designed experiments.

• Describing the benefits of multifactor experiments.

• Differentiating between the impact of a model and the impact of the action taken from that model.

• Fitting incremental response models to evaluate the unique contribution of a marketing message, action, intervention or process change on outcomes.

Optimization Concepts for Data Science

• Identifying and formulating appropriate approaches to solving various linear, nonlinear and efficiency optimization problems.

• Creating optimization models commonly used in industry.

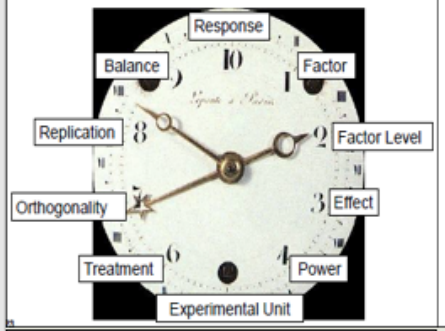
• Formulate optimization problems and how to make their formulations efficient by using index sets and arrays.

• Formulating and solving a data envelopment analysis.

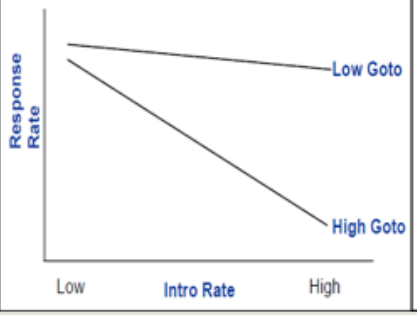
• Solving optimization problems using the OPTMODEL procedure in SAS.

Course demonstrations include examples of data envelopment analysis and portfolio optimization.

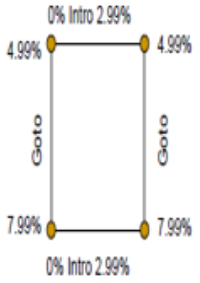
Basic Terms in Design of Experiments (DOE)



Detecting Interactions between Factors



Factorial Arrangement of the Treatments

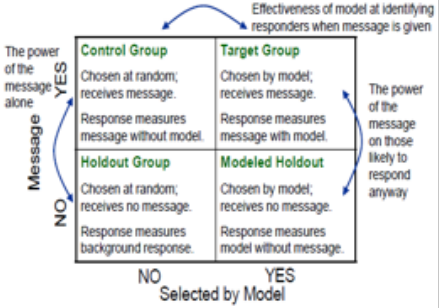


Two-Level Full Factorial Coding

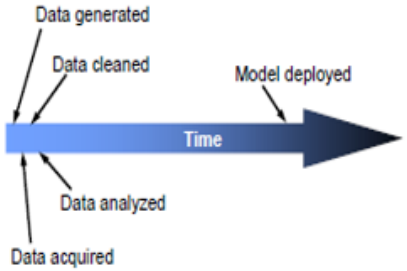
I	A	B	AB
+1	+1	+1	+1
+1	+1	-1	-1
+1	-1	+1	-1
+1	-1	-1	+1

Focus on Designs: A-B Tests

Good Test Design Measures the Impact of Both the Message and the Model



Scoring Pitfalls: Population Drift



Four Types

Persuadables: Respond only with offer

Loyal Customers: Offer irrelevant, Likely to respond

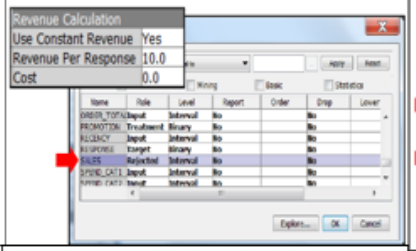
Lost Causes: Offer irrelevant, Unlikely to Respond

Do Not Disturbs: Less likely to respond with offer

RESPONSE:	YES	NO
Offer = YES	Persuadables + Sure things	Do Not Disturbs
Offer = NO	Sure things	Lost Causes + Do Not Disturbs

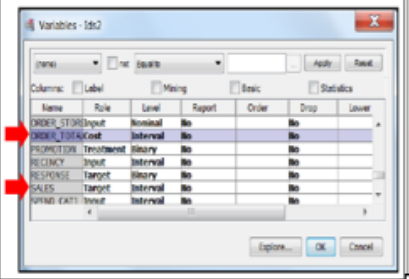
Constant Revenue and Constant Cost

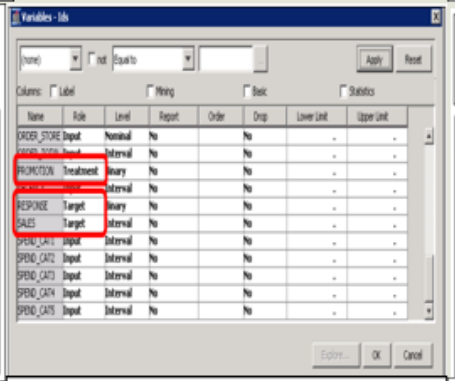
- No use of a cost variable
- No use of an interval target variable (but the property **Use Constant Revenue** would override it)

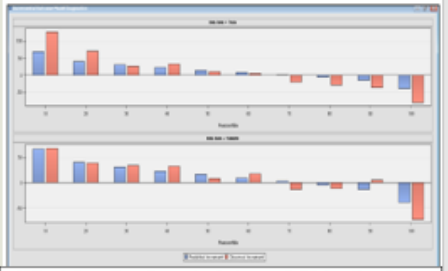


Variable Revenue and Variable Cost

Set the role of the cost variable to **Cost** in the metadata definition.



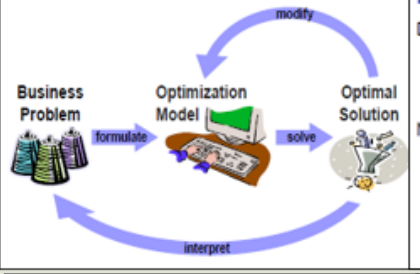




PROC OPTMODEL Programs

```
proc optmodel;
   /* declare se
   /* declare
   solve;
   /* print

quit;
```

The Optimization Modeling Process



Classification of Mathematical Optimization Problems

Discussed in this course:

- **LP:** f(x) and $c_1(x)$, ... ,$c_m(x)$ are linear functions.
- **NLP:** f(x) and $c_1(x)$, ... ,$c_m(x)$ are continuous but not *all* linear functions.

Not discussed in this course:

- ILP: f(x) is linear, x must be integer, and $c_1(x)$, ... ,$c_m(x)$ are linear functions.
- MILP: f(x,y) is linear, x must be integer, and $c_1(x,y)$, ... ,$c_m(x,y)$ are linear functions, but y can be fractional.

Optimization: Going Up or Going Down?

- **LP:** The default settings work almost every time. You only need to push the correct button.
- **ILP/MILP:** Some more difficult problems might require different solver options. What happens when you push the red button?
- **NLP:** Your starting point might determine where you end up. "I've a feeling we're not in Kansas anymore."

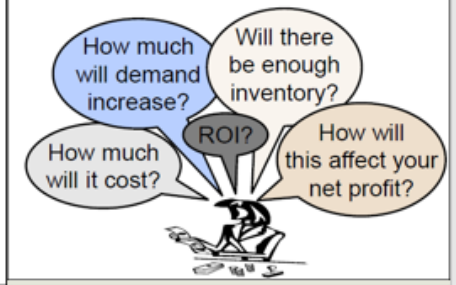
Three-Dimensional Example: Primal Simplex

The following LP has decision variables a, b and c.

Minimize 225a + 117b + 420c

subject to
a + b + 3c >= 12
3a + b +4c >= 19
a >= 0, b >= 0, c >= 0

What Are the Constraints and Objective?



Furniture-Making Problem Data

	Labor (hrs)	Metal (lbs)	Wood (ft²)	Selling Price ($)
Desks	2*Desks	Desks	3*Desks	94*Desks
Chairs	Chairs	Chairs	3*Chairs	79*Chairs
Bookcases	3	1	4	125
Bedframes	2	1	4	109
Cost ($)	14	20	11	
Availability	225	117	420	

- What should the objective be?

Maximize NetProfit = Revenue – Cost

Reading the Product Data: Step by Step

```
read data product_data into PRODUCTS=[Item]
  {r in RESOURCES} <required[Item,r]=col(r)>
  selling_price;
```

required			selling_price
2	1	3	94
1	1	3	79

PRODUCTS

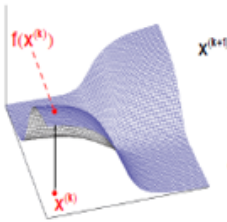
RESOURCES

SAS Data Set: **work.product_data** (first 2 rows)

	Item	Selling_Price	labor	metal	wood
1	desks	94	2	1	3
2	chairs	79	1	1	3

General Strategy for Solving NLPs

Most of the algorithms for solving NLPs are iterative.



$x^{(k+1)} = x^{(k)} + \alpha^{(k)} s^{(k)}$

$x^{(k)}$ iterate

$s^{(k)}$ search direction

$\alpha^{(k)}$ step length

Local information at $x^{(k)}$ is used to find the direction $s^{(k)}$, so essentially f(x) is approximated by a quadratic at $x^{(k)}$.

Two-Dimensional Constrained Example

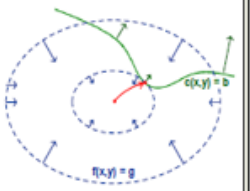
maximize f(x,y) – λ*[c(x,y)–b]

subject to ~~c(x,y) = b~~

The solid line is the set of feasible solutions.

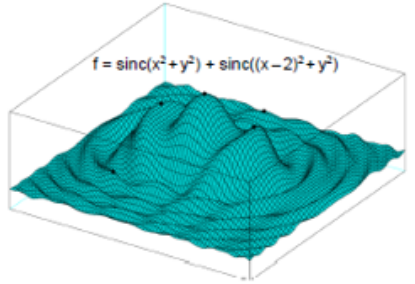
Dashed lines are contours of the objective function.

Arrows are gradients of the objective function and constraint.



Smooth Does Not Mean Well-Behaved

The optimality conditions can be satisfied at *local* optima.



Example: Portfolio Optimization Problem

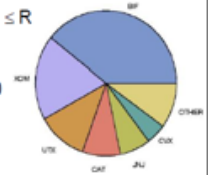
Select a portfolio that maximizes expected return while not exceeding a specified risk R (measured by variance).

maximize $\sum_{j=1}^{n} r_j x_j$

subject to $\sum_{i=1}^{n}\sum_{j=1}^{n} x_i \sigma_{ij} x_j \le R$

$\sum_{j=1}^{n} x_j = 1$

$x_1 \ge 0, \ldots, x_n \ge 0$



x_j proportion of stock j

r_j expected return of stock j

σ_{ij} covariance between stocks i and j

Thank you!