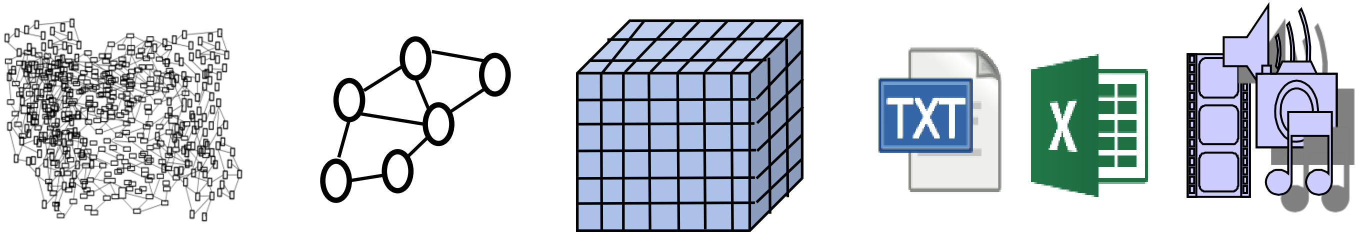
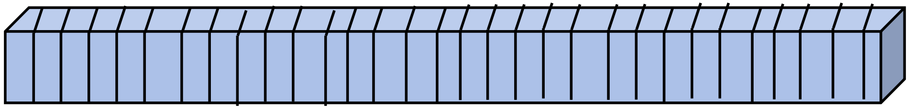
BIG DATA ANALYTICS STRUCTURES





Presented by

David Haertzen

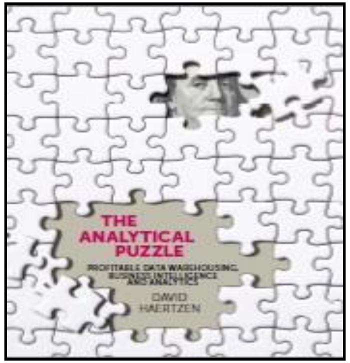
First Place Learning

About the Instructor

- Author of the book, The Analytical Puzzle, plus numerous articles and courses
- Trail blazer and thinker
- Provided services to organizations such as: Allianz Life, 3M, Mayo Clinic, IBM, Fluor Daniel, Procter & Gamble and Synchrono – from start up to multinational
- Thought provoking presenter in the areas of:
  - Profitable Analytics
  - Data Modeling
  - Data Warehousing
  - Enterprise Architecture
  - Business Intelligence
  - SQL
- University of Minnesota
  MBA, University of St Thomas
- Home Page: http://davidhaertzen.com

David Haertzen

Author and Instructor



Session Structure

Module 1: Overview of Analytic Applications

- Waves of Analytic Applications
- Analytic Methodology
- Analytic Architecture
- Data Structures

Module 2: Flat Data for Predictive Analytics

- Example Predictions
- Data for Predictive Analytics
- Developing Predictive Models

Module 3: Unstructured and Semi-structured Data

- Unstructured and Semi-structured Data
- Text Mining
- Image Mining

Objectives:

- Understand what Analytics is, its goals, and its components
- Learn new Analytics terms
- Be able to select the right data structure to match the analytic problem
- Understand how to benefit from Analytics
- Be able to discuss Analytical methods and tools
- Be prepared to learn more about Analytics

BIG DATA ANALYTICAL STRUCTURES

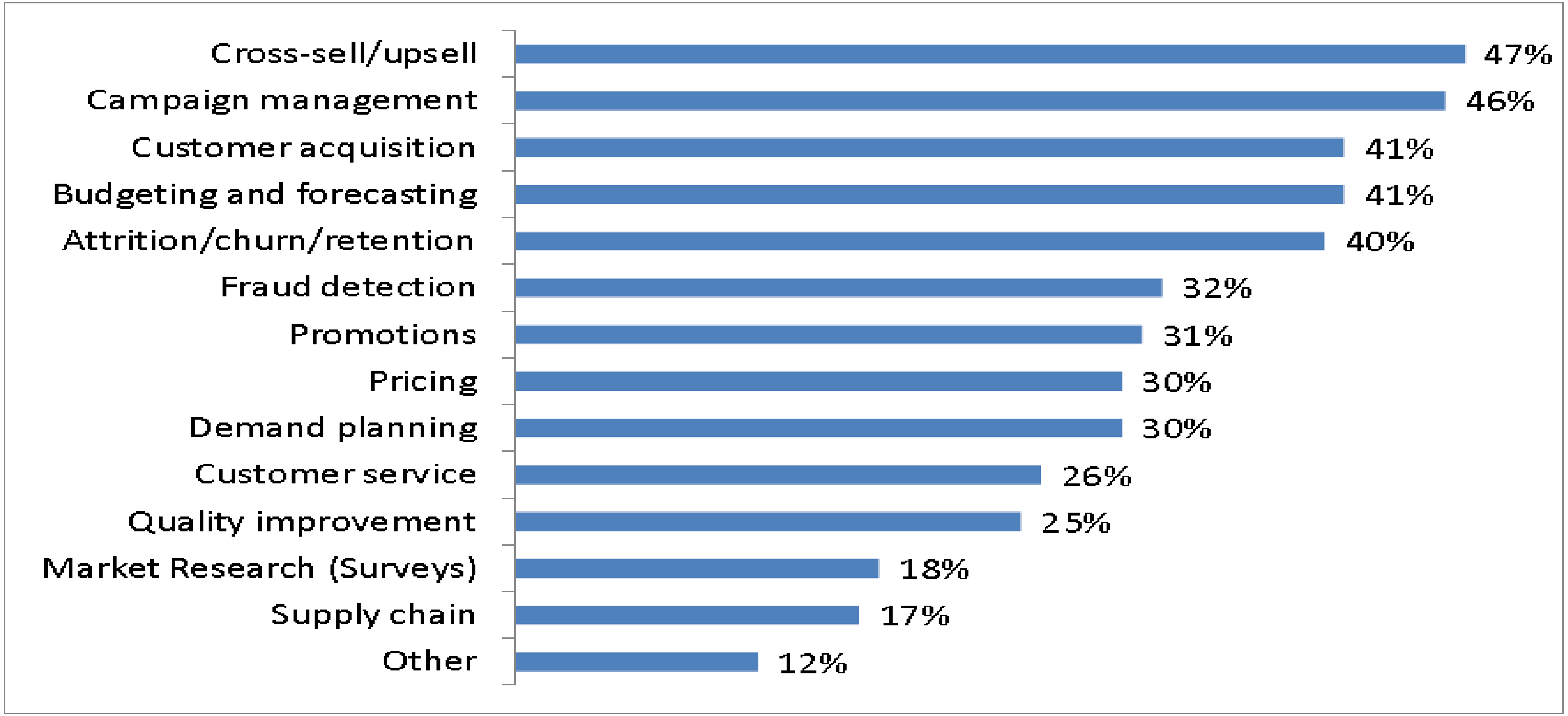
Topic I:

Overview of Analytical Applications

I. Overview of Analytic Applications

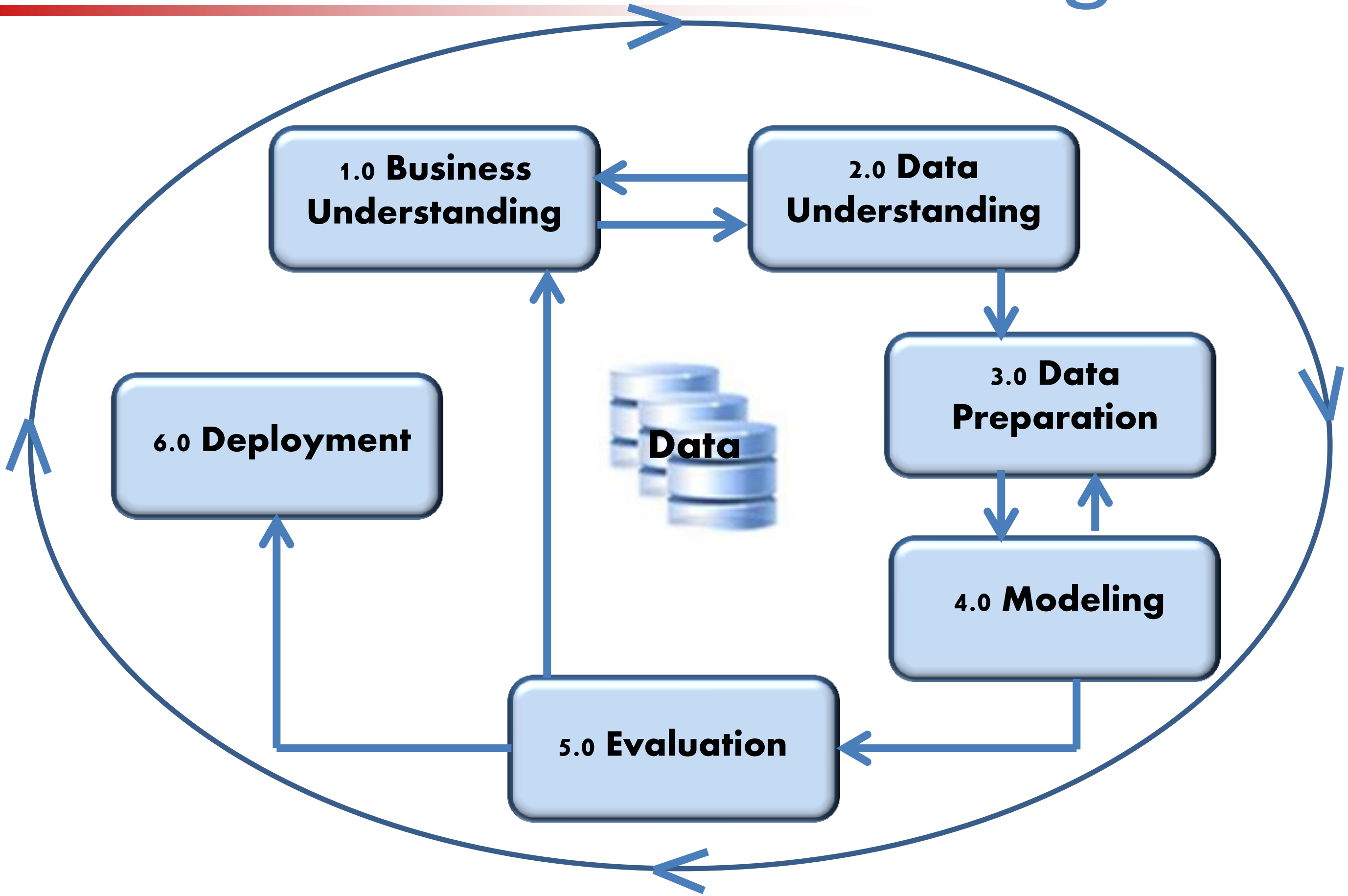
- Advanced Analytical Applications
- Analytic Methodology
- Analytic Architecture

Advanced Analytics Applications



From Wayne Eckerson, "Predictive Analytics: Extending the Value of Your Data Warehousing Investment," TDWI, 2007. Based on 166 respondents that had implemented predictive analytics.

CRISP-DM Data Mining Methodology



CRISP-DM (CRoss Industry Standard Process for Data Mining)

A methodology developed in late 1990s, partially funded by the European Commission under the ESPRIT Program.

Contributors:

- NCR Systems Engineering
- SPSS Inc.
- DaimlerChrysler
- OHRA Verzekering en Bankk Groep

Big Data

Big Data is data that has such large volume, great variety or rapid velocity that it cannot be effectively managed using tradition relational databases.

Volume	Big data may require huge volumes of data storage – in the tens of terabytes to petabytes and zetabytes.
Variety	Big data often comes in a variety of formats which can change depending on the source and application.
Velocity	Big data often arrives faster than traditional single computer systems can handle, such as information from sensors or the Internet.

Analytics Architecture Components

Data Sources

- Data Brokers
- Social Media
- Operational Systems
- Text Files
- Images
- Logs
- Government
- Internet of Things
- Spreadsheets

More ➢

Logical Data Lake

- Operational Systems
- EDW
- Data Marts
- Hadoop
- NoSQL
- Sand Boxes
- tbd
- Cloud
- In Memory

More ➢

Analytical Models

- Neural Network
- Regression
- Decision Tree
- Machine Learning
- Cluster
- Montecarlo
- Optimization
- Simulation
- Forecast

More ➢

Uses

- Report
- Dashboard
- Scorecard
- Recommend Engine
- Anti-fraud
- Campaign Management
- Algorithmic Trading
- Customer Intelligence
- Location Analysis

More ➢

Increasing Analytics Performance

<u>Database Technology – SQL and NoSQL:</u>
- ✓ SQL Traditional Database
- ✓ Data Warehouse Appliance
- ✓ Columnar Database
- ✓ In Memory Database
- ✓ OLAP / Cube Database
- ✓ NoSQL Databases

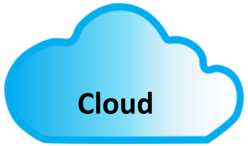
<u>Scale It Up:</u>
- ✓ *Memory*
- ✓ *Flash / SSD*
- ✓ *CPUs and Cores*
- ✓ *Dedicated Fast Disks*

<u>Scale It Out:</u>
- ✓ Grid - Data Synapse
- ✓ In memory Grid - Apache Ignite, others
- ✓ DIY Grid
- ✓ GPU - CUDA, OpenCL, BOINC
- ✓ Supercomputer / Minisuper Computer
- ✓ Distributed Data & Calcs – Hadoop, Spark
- ✓ Streaming – Spark and Storm
- ✓ Cloud

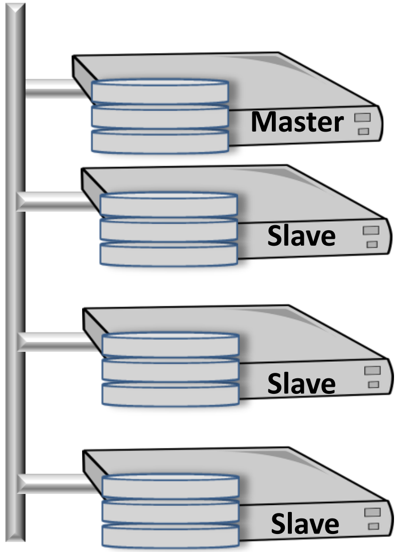
<u>Improve Design and Implementation:</u>
- ✓ *Buy Pre-analyzed and Aggregated Data*
- ✓ *Dimension Reduction*
- ✓ *Sample Data*
- ✓ *Faster Algorithms*
- ✓ *Data Filter*
- ✓ *Data Vault*
- ✓ *Indexing*
- ✓ *Query Optimization*
- ✓ *Change Data Capture / Streaming*

Big Data Scale Out Architecture

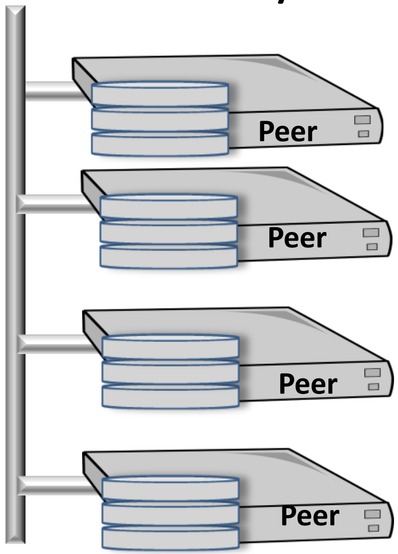
Scale Out
High Performance Computing



Master Save Grid / Cluster



Peer Grid / Cluster

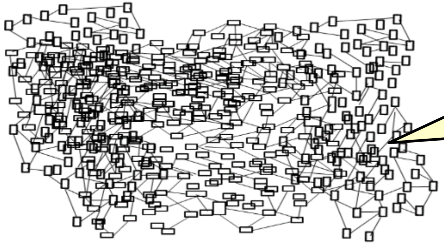


<u>Shared Nothing</u>
Each node has its own
OS, Compute and Storage

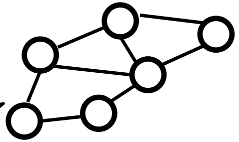
Big Data / Sample Data

	Plus (+)	Minus (-)
Big Data is data that is so voluminous that it cannot be managed using traditional databases such as relational databases. This data is often unstructured and consists of text, images, video, and audio.	➢ May reveal outliers and exception opportunities ➢ May reveal new trends ➢ Analysis of unstructured data requires big data ➢ Analysis of physical processes requires big data ➢ More accurate than a sample	➢ Requires more time to gather ➢ Costs more to store ➢ Takes longer to analyze ➢ Doesn't fit into memory
Sample Data is a portion of a population selected for statistical analysis and predictive analytics.	➢ Costs less to gather ➢ Costs less to store ➢ Can have a high confidence level ➢ Compatible with predictive analytics algorithms ➢ Faster results can be quickly applied	➢ Requires discrete facts ➢ May miss outliers and exceptions

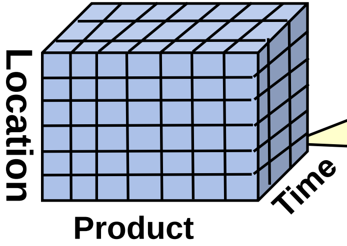
Analytical Data Structures



Normalized Structured Data Enable Transaction Processing:
- Create, Read, Update and Delete



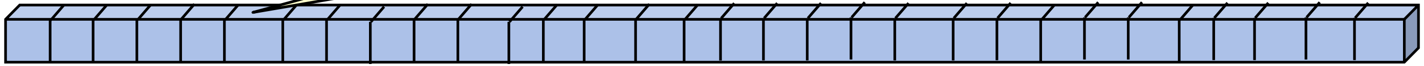
Graphs Enable Scheduling and Affinity Analysis:
- Critical Path, Logistics, Factories, Work Flows, MBA. Social Media



Cubes and Star Schemas Enable Human Analytic / Visual Data Exploration:
- Roll Up, Drill Down, Slice and Dice

Unstructured and Semi-Structured Data:
- Text, spreadsheets, XML, images, video, audio, social media, machine sensor

Flattened Data Supports Analytical Algorithms:
- Regression, Decision Tree, Neural Net, MBA



BIG DATA ANALYTICAL STRUCTURES

Topic II:

Flat Data for Predictive Analytics

II. Flat Data for Predictive Analytics

- Example Predictions
- Data for Predictive Analytics
- Developing Predictive Models

Predicting Who is Most Like to …

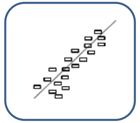
Behave Well:

- Sell Successfully
- Buy a Product
- Buy a Premium Product
- Respond to a Treatment
- Respond to a Campaign
- Achieve High Grades
- Finish College in 4 Years
- Stay Loyal for Years
- Drive Safely
- Recommend Product
- Use Self-Service

Behave Badly:

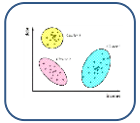
- Commit a Violent Crime
- Commit Fraud
- Switch to a Competitor
- Cancel an Order
- Drop Out from School
- Skip Bail
- Return a Product
- Quit a Job
- Make a Terrorist Attack
- Have an Automobile Accident
- Make Expensive Requests

Predictive Analytical Models



Regression

A statistical method that predicts the value of one variable based on the value(s) of one or more numeric variables. For example, the variable of wine price might be predicted based on winter rainfall, average growing season temperature and harvest rainfall.

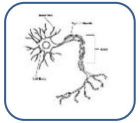


Cluster

An analytic method based on grouping data points with a large degree of affinity. The data points have much in common with data points in the cluster and differ from data points in other clusters.

Decision Tree

A structure that enables large collections of inputs to be classified into homogeneous groups through a series of choices called nodes. The tree is processed from left to right or top to bottom, with the first node called the root node, nodes secondary to the root node called child nodes, and nodes at the bottom called leaf nodes.



Neural Network

A **flexible predictive analytics tool that mimics the learning of the human brain. A neural net model accepts a large collection of known inputs and produces an output that may be continuous-valued. Neural nets include machine learning and so can improve with use.**

Input Data Set Roles

1. **Training Set** — The Training Set is used to build the initial analytical model. Algorithms are developed that predict the Training Set.

2. **Test Set** — The Test Set is used to refine the analytical model developed using the Training Set. The model is refined to predict the Test Set also.

3. **Evaluation Set** — The Evaluation Set is used to judge the effectiveness of the analytical model refined using the Test Set.

4. **Score Set** — The Score Set is the deployment version of the data used to produce actual predictions.

Analytical Models Require Flat Inputs

Training Data Set each row or record is processed separately and contains the input attributes needed to make a prediction or classification.

Identifier

Attributes

Target (What is to be predicted)

	A	B	C	D	E	F	G	H	I	J	K
1	ClientNbr	Tenure	ZipCode	Age	Gender	AddrChangeCount	PhoneChangeCount	EmailChangeCount	RecentTxnCount	RecentTxnTotalAmt	FraudScore
2	1001	1	55124	25	M	2	2	2	8	50000	100
3	1002	5	55123	25	F	1	1	0	1	5000	20
4	1003	5	55123	55	F	1	0	0	1	5000	20
5	1004	2	55125	35	M	1	0	0	0	0	80
6	1005	1	55124	75	F	2	2	2	7	60000	100
7	1006	6	55313	75	F	0	1	0	1	5000	10

Identifier — Entity identifiers such as customer number are ignored by analytical models.

Attribute — Input value

Target — The predicted or classification value.

Inputs for Predictive Analytics: Data Types

Identifier	Entity identifiers such as customer number are not used as input to analytical models.
Qualitative	Qualitative attributes are descriptive or categorical, rather than numeric. Mathematical operations do not apply to qualitative attributes. **Nominal** attributes are descriptors whose values imply no order, while **ordinal** attributes have order.
Binary	A two valued data element: yes/no, true/false, 1/0. This is a very useful for categorization.
Quantitative	Attributes that are numeric and subject to mathematical operations. **Interval** quantitative attributes lack true zero such as credit scores and time of day. **Ratio** attributes have true zero such as counts, weights and time durations.
Dates/Time	Dates and times (interval attributes) are not readily supported by analytical algorithms. Change to tenure to allow categorization or mathematical operations.
String	A text value such as a person's name or street address: "Sandy Shores" or "PO Box 156". Predictive models do not do well with this type of data.

Binning

Binning is a method that converts open-ended / noisy data into discrete data with a limited number of values. These values can better handled by some analytical techniques such as decision trees that work well with categorical values or discrete / smoothed values.

Binning Techniques:

- Sort data and divide into bins
- Assign ordinal numbers
- Smooth by mean, medians or boundaries

Binning Examples:

- Age bins:
  - 15 = Under 21
  - 25 = 20-29
  - 35 = 30 – 39
  - 45 = 40 – 49
  - 55 = 50 – 59
  - 65 = 60 – 69
  - 75 = Over 69
- Education bins:

  0 = No high school
  1 = High School
  2 = 2 Years College
  3 = 4 Years College
  4 = Masters
  5 = Doctorate
  6 = Post Doctorate

Dimension Reduction

Dimension Reduction is the process of simplifying input factors to predictive analytics algorithms to reduce the number and/or complexity. The process may reduce 100s of factors to a handful.

Methods:

- Drop Missing Values
- Drop Low Variance
- High Correlation
- Backward Feature Elimination
- Factor Analysis
- Principal Component Analysis (PCA)

Benefits:

- Calculations are faster.
- Storage space needed is reduced.
- Models are easier to explain.
- Model is easier to productionize.

Examples:

Wine Price depends on:

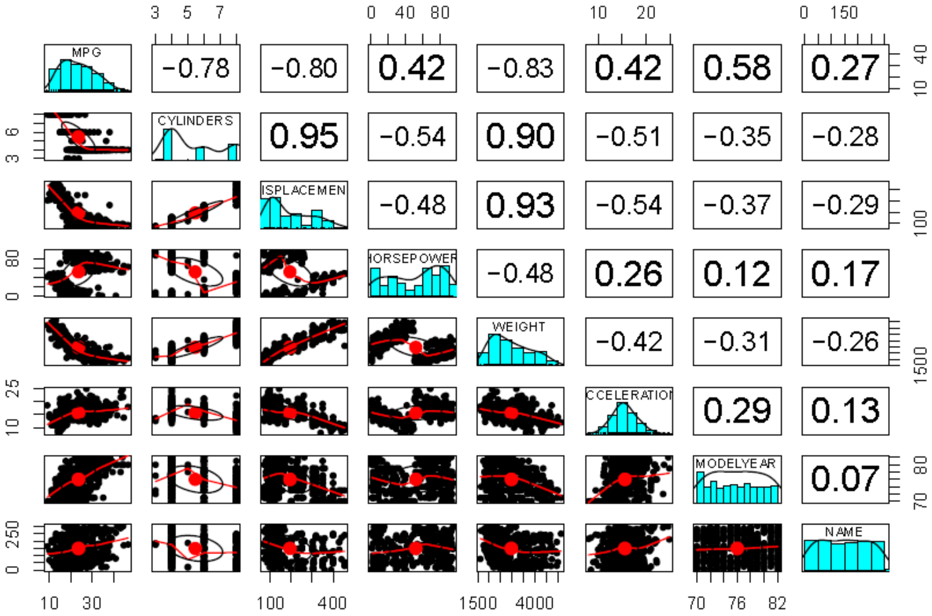
- Winter Rainfall
- Average Growing Season Temperature
- Harvest Rainfall

UPS Route Safety depends on Left Turns

Dimension Reduction
http://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+Vidhya%29

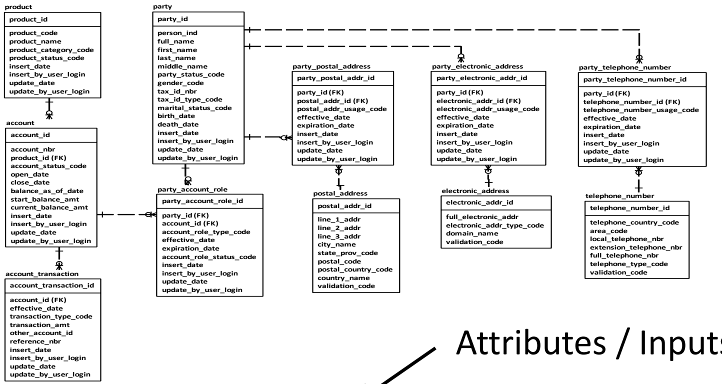
Correlation Matrix

A Correlation Matrix is a table which shows the degree that attributes occur in proportion to each other. Negative numbers show an inverse relationship. This information is used for dimension reduction.



Flatten Relational Structures for Analytics

Relational Data Must Be Flattened through preprocessing steps such as adding the number of customer address changes or financial transactions.



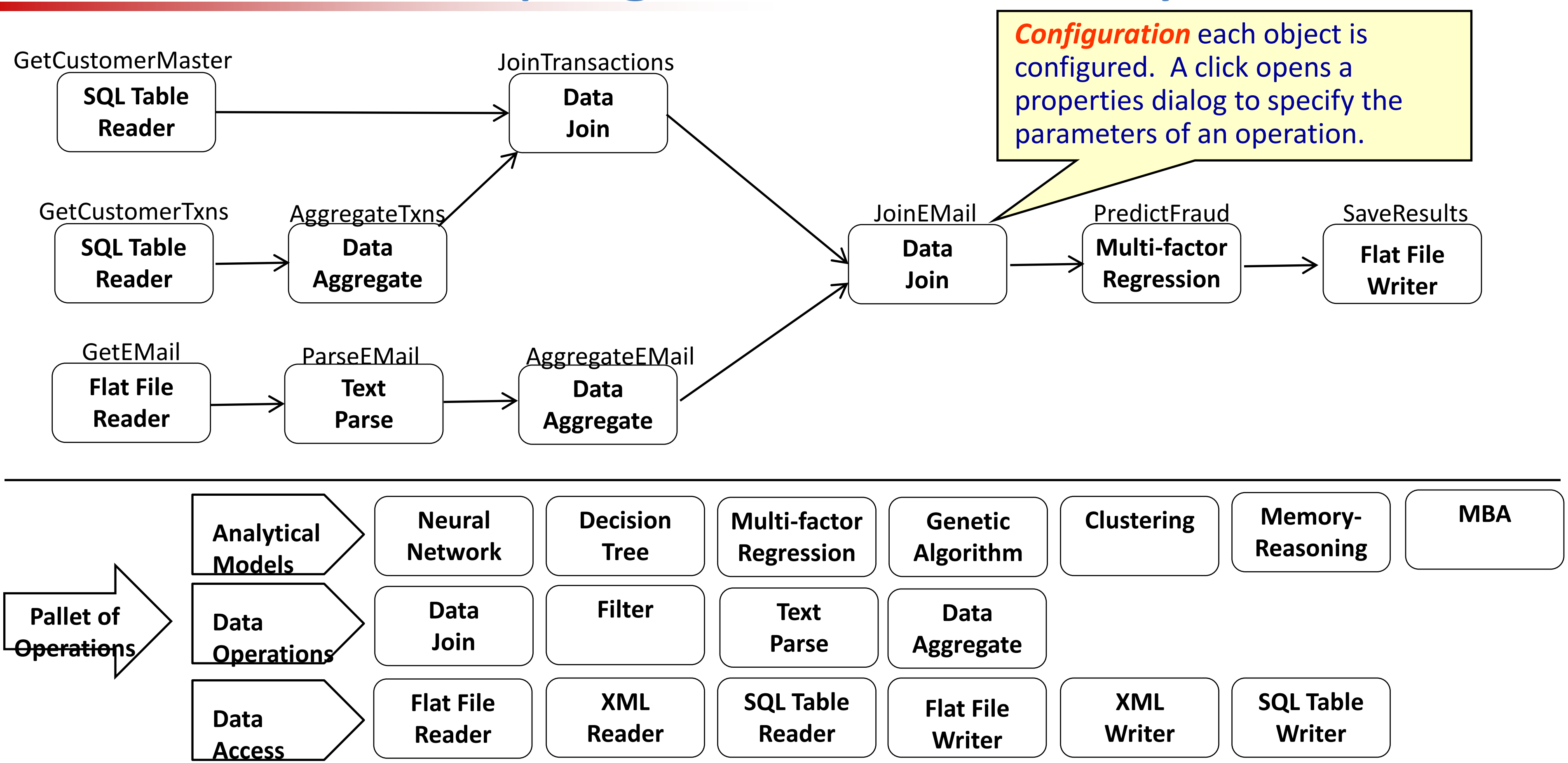
Identifier

Attributes / Inputs

Target
(What is to be
predicted)

	A	B	C	D	E	F	G	H	I	J	K
1	ClientNbr	Tenure	ZipCode	Age	Gender	AddrChangeCount	PhoneChangeCount	EmailChangeCount	RecentTxnCount	RecentTxnTotalAmt	FraudScore
2	1001	1	55124	25	M	2	2	2	8	50000	100
3	1002	5	55123	25	F	1	1	0	1	5000	20
4	1003	5	55123	55	F	1	0	0	1	5000	20
5	1004	2	55125	35	M	1	0	0	0	0	80
6	1005	1	55124	75	F	2	2	2	7	60000	100
7	1006	6	55313	75	F	0	1	0	1	5000	10
8	1007	8	55313	75	M	0	0	1	1	5000	10
9	1008	2	55321	65	M	0	0	1	1	5000	0
10	1009	5	55416	45	M	2	2	2	7	60000	90
11	1010	8	55321	65	F	0	0	0	1	7000	10
12	1011	7	55128	65	F	0	0	0	2	5000	10
13	1012	2	55128	55	M	0	0	0	0	0	10
14	1013	1	55416	65	F	2	2	2	7	60000	80
15	1014	1	55345	65	M	2	2	2	7	60000	90
16	1015	1	55340	75	M	1	0	0	0	0	20
17	1016	1	55341	75	F	0	0	0	2	6000	20
18	1017	1	55782	75	F	1	0	0	2	6000	20
19	1018	1	55340	65	M	2	2	2	7	60000	60
20	1019	3	55124	65	F	0	0	0	1	5000	20
21	1020	3	55124	55	M	1	0	0	1	5000	20

Developing Predictive Analytics



FIRST PLACE LEARNING
GOAL DRIVEN TRAINING

Assessing Numeric Predictions

Assessing Numeric Predictions includes a determination of the accuracy of the prediction compared to actual.

- Descriptive statistics – standard deviation, variance, etc.
- Quantify cases where prediction is inside and outside business tolerance.

Assessing Classification Models

A Confusion Matrix is a method for evaluating Classification Models that quantifies the number and proportion of correct and incorrect classifications through use of a table.

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
- **Accuracy** = (TN+TP)/n (60 + 105) / 190 = 87%
- **Error Rate** = (FN+FP)/n (10 + 15) / 190 = 13%

n=190	Predicted: No	Predicted: Yes	
Actual: No	TN=60	FP=15	75
Actual: Yes	FN=10	TP=105	115
	70	120	

BIG DATA ANALYTICAL STRUCTURES

Topic III:

Unstructured and Semi-structured Data

III. Unstructured and Semi-structured Data

- Unstructured Data
- Text Mining
- Image Mining
- Semi-structured Data

Unstructured Data

Unstructured Data is data that does not have a defined data model or format such as: text, images and sounds. It has been estimated that 70 to 90 percent of all data is unstructured. Analysts are working to organize unstructured data into structured data.

Free format text

Video

Audio

Spreadsheets

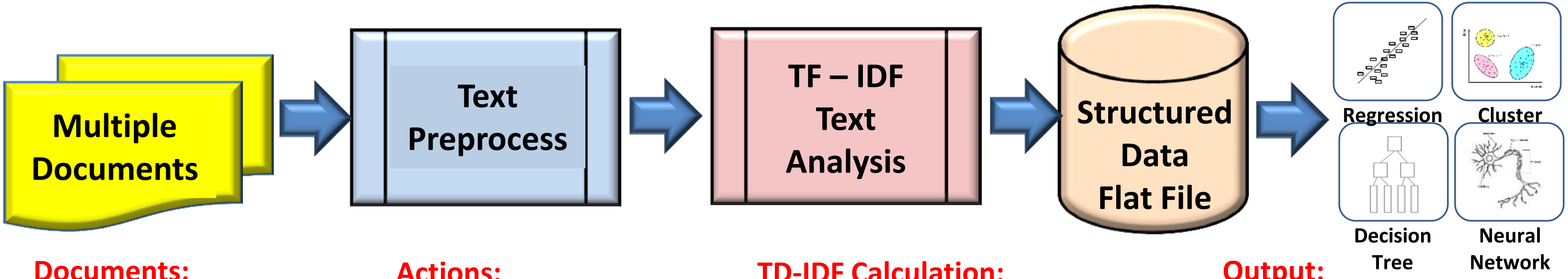
Still images

Music

Machine generated
Sensor output

Text Mining Using TF-IDF

Text Frequency – Inverse Data Frequency is a method of quantifying the strength of words that make up documents – based on relative frequency of words.



Documents:
- Emails
- Reports
- Tweets
- Comments and Notes
- Separate Files
- NoSQL Database Fields

Actions:
- Bag of Words / Tokenize
- Remove punctuation
- Make lower case
- Remove stop words
- Remove local words
- Resolve to word stems

TD-IDF Calculation:
- Text Frequency (TF) for each word in a document = specific word count / total words in document count
- Inverse Document Frequency (IDF) = log e (total # docs / total docs with word)
- TF-IDF = TF * IDF

Output:
- One record / document
- Numbers not words

Document_Id	Accident	Automobile	Fender	Injury	Police	…	XRay
123001	0	0	0	.13	.20		.05

Knime Text Processing

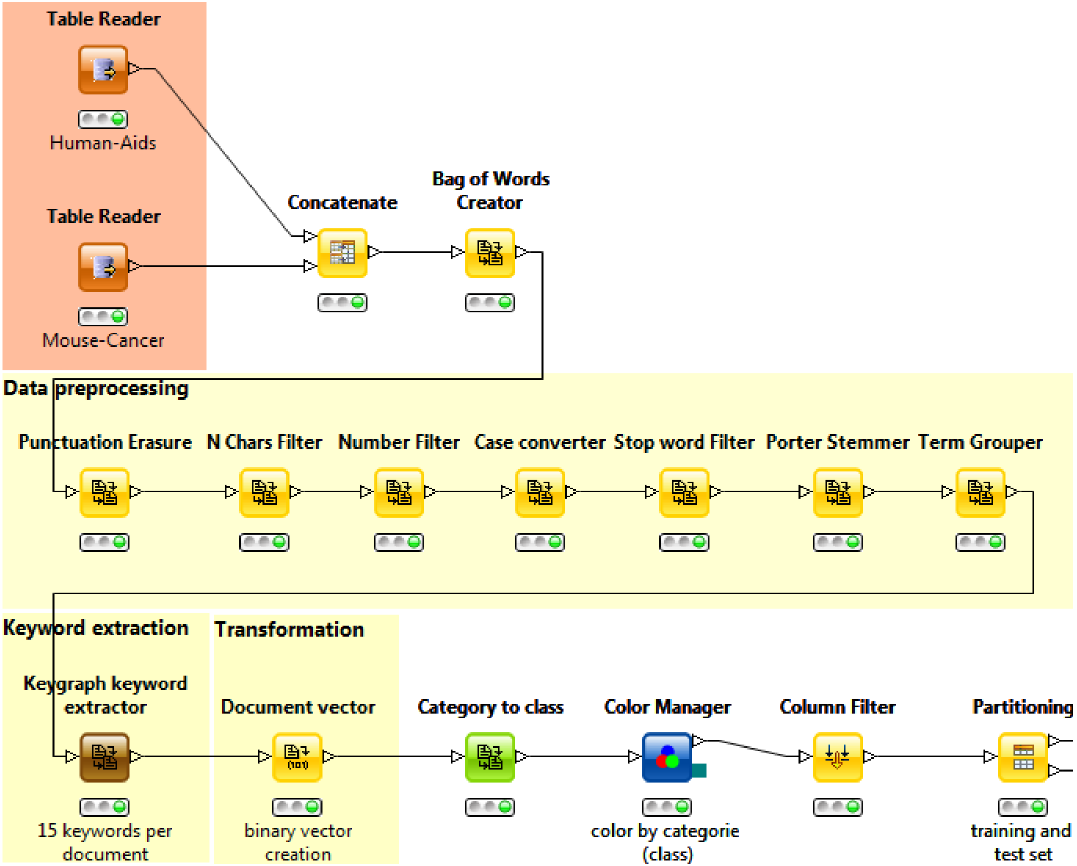


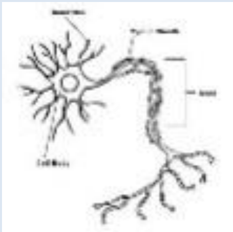
Image Mining / Google Example

Image Mining is the process of extracting meaning from image data.

Unstructured Data

Google Street View

Neural Net Trained
With
200,000 Images of
Street Numbers



Structured Data

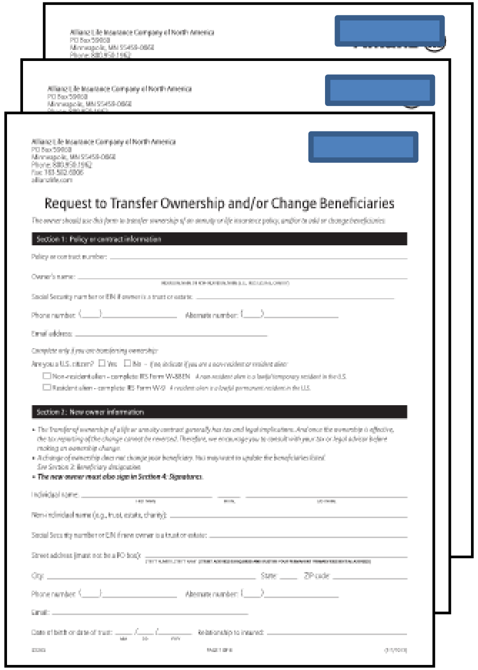
… 379 Rue de Napoli …
… 61 Hudson Street …
… 972 Seventh Avenue …
… 11 East Main Street …
… 6624 Cleveland Road …
… 98 Wilshire Blvd …
… 66 Whipple Road …
… 175 Duncan Drive …
… 2 Waldo Way …
… 205 Donald Lane …
… 100 Industrial Blvd …

http://www.technologyreview.com/view/523326/how-google-cracked-house-number-identification-in-street-view/

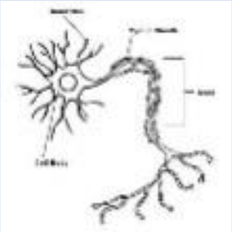
Image Mining / Insurance Example

Image Mining can be used to improve the speed and accuracy of document input.

Insurance Form Images



Neural Net Trained with text recognizer pulls 99.9% plus data correctly – including handwriting



Structured Data

Request = Change of Bene
Date = 2015-09-15
Policy = 1234567
….

Music Recognition Example

Music Recognition can be used to identify and classify music.

Music

Structured Data

Title = Hey Jude
Artist = Beatles
…

Teaching Patterns to First Graders

First Graders Learn Patterns using App

Smart Phones learn to recognize Objects using Deep Learning

http://bridgingapps.org/2012/08/bridgingapps-reviewed-app-a-1st-grade-pattern-recognition-game-for-ipad/

http://www.purdue.edu/newsroom/releases/2014/Q1/smartphone-to-become-smarter-with-deep-learning-innovation.html

Quantifying the Face

Facial Action Coding System (FACS)

AU Number	FACS Name
0	Neutral face
1	Inner Brow Raiser
2	Outer Brow Raiser
4	Brow Lowerer
5	Upper Lid Raiser
6	Cheek Raiser
7	Lid Tightener
8	Lips Toward Each Other
9	Nose Wrinkler
10	Upper Lip Raiser
11	Nasolabial Deepener

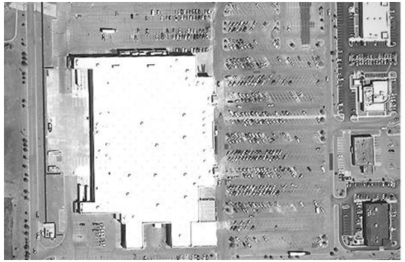
AU Number	FACS Name
40	Sniff
50	Speech
80	Swallow
81	Chewing
82	Shoulder shrug
84	Head shake back and forth
85	Head nod up and down
91	Flash
92	Partial flash
97*	Shiver/Tremble
98*	Fast up-down look

Observation_Id	Code_06	Code_07	Code_08	Code_09	Code_10	…	Code_98
123001	0	0	0	3	5		3
123002	5	0	0	1	2		0

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

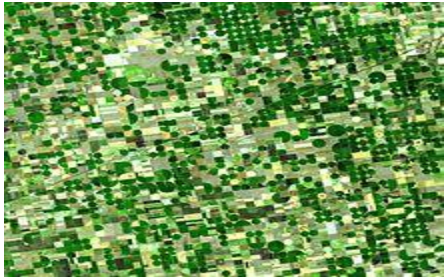
Satellite and Aerial Imaging

Retail



Investors use satellite images of retail parking lots to predict same store sales. Samples 100 stores. Counts cars.

Agriculture

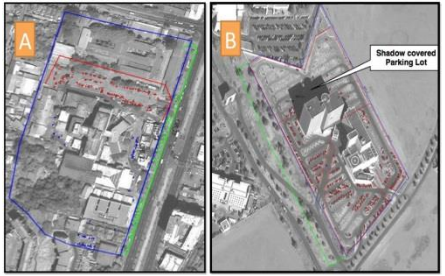


Farmers monitor their individual crops. Investors, government, insurance and others analyze trends.

Mining

Mining companies analyze their operations. Investors, ecology activists, governments and others analyze trends.

Health



Disease control analysts use images of hospital parking lots to gain an indication of infectious disease outbreaks.

Semi-structured Data

Semi-structured Data is data that is self-described using tags such as XML and JSON.

XML

Extensible Markup Language (**XML**) is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable. It is defined by the W3C's XML 1.0 Specification[2] and by several other related specifications,[3] all of which are free open standards.[4] https://en.wikipedia.org/wiki/XML

```
<Product>
   <ProductNbr>123987</ProductNbr>
   <ProductName>Widget</ProductName>
</Product>
```

JSON

JSON, (canonically pronounced /ˈdʒeɪsən/ ***JAY-sən***;[1] sometimes **JavaScript Object Notation**), is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs. It is the primary data format used for asynchronous browser/server communication (AJAJ), largely replacing XML (used by AJAX). https://en.wikipedia.org/wiki/JSON

```
{"Product":[
   {"ProductNbr":"123987", "ProductName":"Widget"}
]}
```

Analytical Models Require Flat Inputs

Training Data Set each row or record is processed separately and contains the input attributes needed to make a prediction or classification.

Identifier

Attributes

Target (What is to be predicted)

	A	B	C	D	E	F	G	H	I	J	K
1	ClientNbr	Tenure	ZipCode	Age	Gender	AddrChangeCount	PhoneChangeCount	EmailChangeCount	RecentTxnCount	RecentTxnTotalAmt	FraudScore
2	1001	1	55124	25	M	2	2	2	8	50000	100
3	1002	5	55123	25	F	1	1	0	1	5000	20
4	1003	5	55123	55	F	1	0	0	1	5000	20
5	1004	2	55125	35	M	1	0	0	0	0	80
6	1005	1	55124	75	F	2	2	2	7	60000	100
7	1006	6	55313	75	F	0	1	0	1	5000	10

Identifier — Entity identifiers such as customer number are ignored by analytical models.

Attribute — Input value

Target — The predicted or classification value.

Session Review

Module 1: Overview of Analytic Applications

- Waves of Analytic Applications
- Analytic Methodology
- Analytic Architecture
- Data Structures

Module 2: Flat Data for Predictive Analytics

- Example Predictions
- Data for Predictive Analytics
- Developing Predictive Models

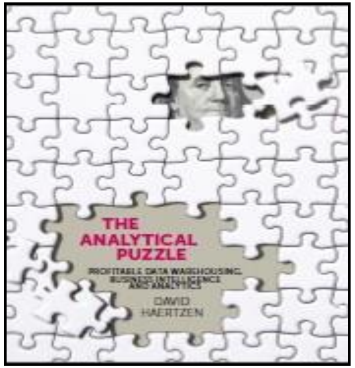
Module 3: Unstructured and Semi-structured Data

- Unstructured and Semi-structured Data
- Text Mining
- Image Mining

David Haertzen – Contact Information

David Haertzen

Author and Instructor



http://www.davidhaertzen.com/

http://www.linkedin.com/davidhaertzen

http://ecm.elearningcurve.com/David_Haertzen_s/89.htm

Twitter: #BigHeart7

david at davidhaertzen dotCom