Chapter 2

Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization

Learning Objectives (1 of 2)

2.1 Understand the nature of data as it relates to business intelligence (BI) and analytics

2.2 Learn the methods used to make real-world data analytics ready

2.3 Describe statistical modeling and its relationship to business analytics

2.4 Learn about descriptive and inferential statistics

2.5 Define business reporting, and understand its historical evolution

Learning Objectives (2 of 2)

2.6 Understand the importance of data/information visualization

2.7 Learn different types of visualization techniques

2.8 Appreciate the value that visual analytics brings to business analytics

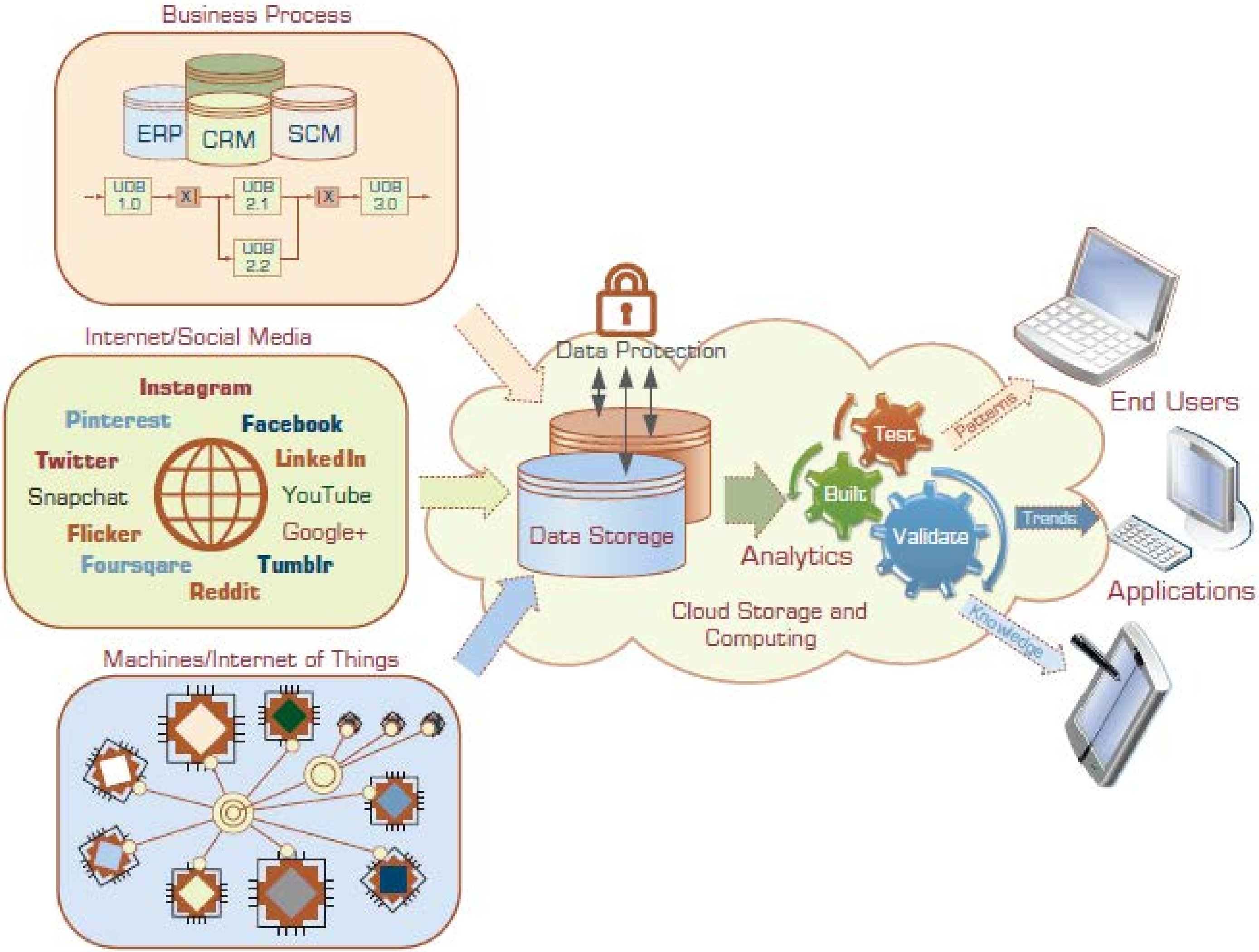
2.9 Know the capabilities and limitations of dashboards

Why do we need to understand the nature of data?

The Nature of Data (1 of 2)

- Data: a collection of facts
  - usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, images, …
- Data is the lowest level of abstraction (from which information and knowledge are derived)
- Data is the source for information and knowledge
- Data quality and data integrity → critical to analytics

The Nature of Data (2 of 2)



Metrics for “Analytics Ready” Data

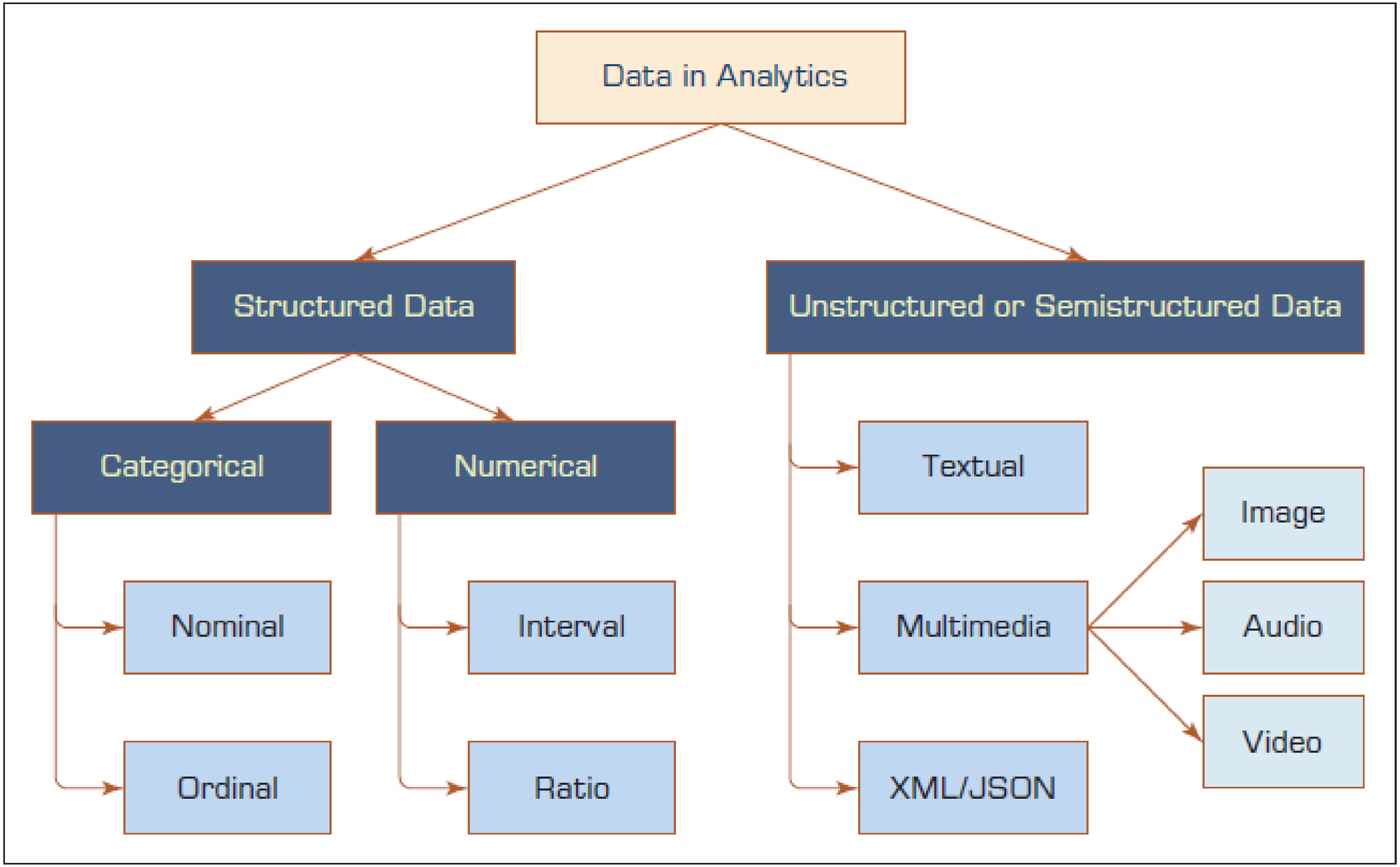
- Data source reliability
- Data content accuracy
- Data accessibility
- Data security and data privacy
- Data richness
- Data consistency
- Data currency/data timeliness
- Data granularity
- Data validity and data relevancy

Structured VS Unstructured DATA

A Simple Taxonomy of Data (1 of 2)

- Data (datum—singular form of data): facts
- Structured data
  - Targeted for computers to process
  - Numeric versus nominal
- Unstructured/textual data
  - Targeted for humans to process/digest
- Semi-structured data?
  - XML, HTML, Log files, etc.
- Data taxonomy…

A Simple Taxonomy of Data (2 of 2)



Application Case 2.1

Medical Device Company Ensures Product Quality While Saving Money

Questions for Discussion

1. What were the main challenges for the medical device company? Were they market or technology driven?
2. What was the proposed solution?
3. What were the results? What do you think was the real return on investment (ROI)?

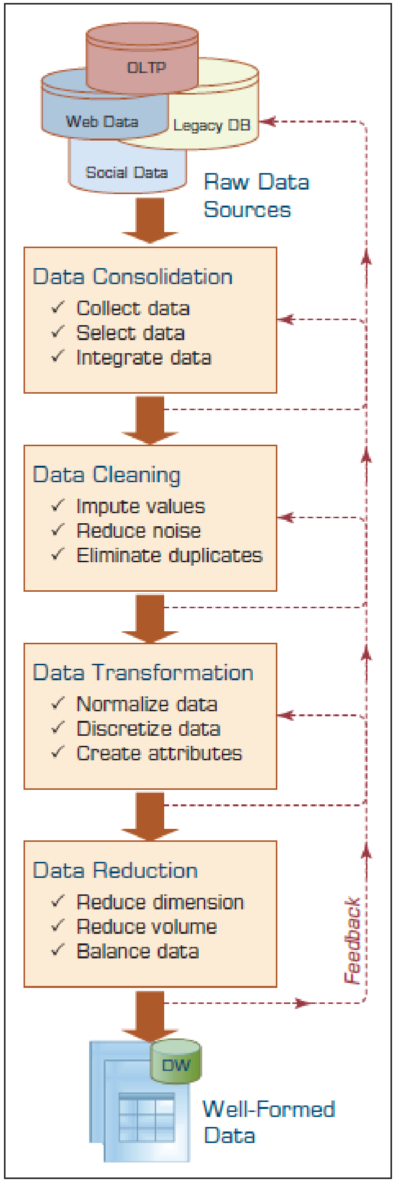
The Art and Science of Data Preprocessing (1 of 2)

- The real-world data is dirty, misaligned, overly complex, and inaccurate
  - Not ready for analytics!
- Readying the data for analytics is needed
  - Data preprocessing
    - Data consolidation
    - Data cleaning
    - Data transformation
    - Data reduction
- Art – it develops and improves with experience

The Art and Science of Data Preprocessing (2 of 2)

- Data reduction

1. Variables
   - Dimensional reduction
   - Variable selection
2. Cases/samples
   - Sampling
   - Balancing / stratification



Data Preprocessing Tasks and Methods (1 of 3)

Table 2.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	**Subtasks**	**Popular Methods**
Data consolidation	Access and collect the data Select and filter the data Integrate and unify the data	SQL queries, software agents, Web services. Domain expertise, SQL queries, statistical tests. SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as “ML”; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.

Data Preprocessing Tasks and Methods (2 of 3)

Main Task	Subtasks	Popular Methods
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or -1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range-or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.

Data Preprocessing Tasks and Methods (3 of 3)

Main Task	Subtasks	Popular Methods
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Application Case 2.2 (1 of 6)

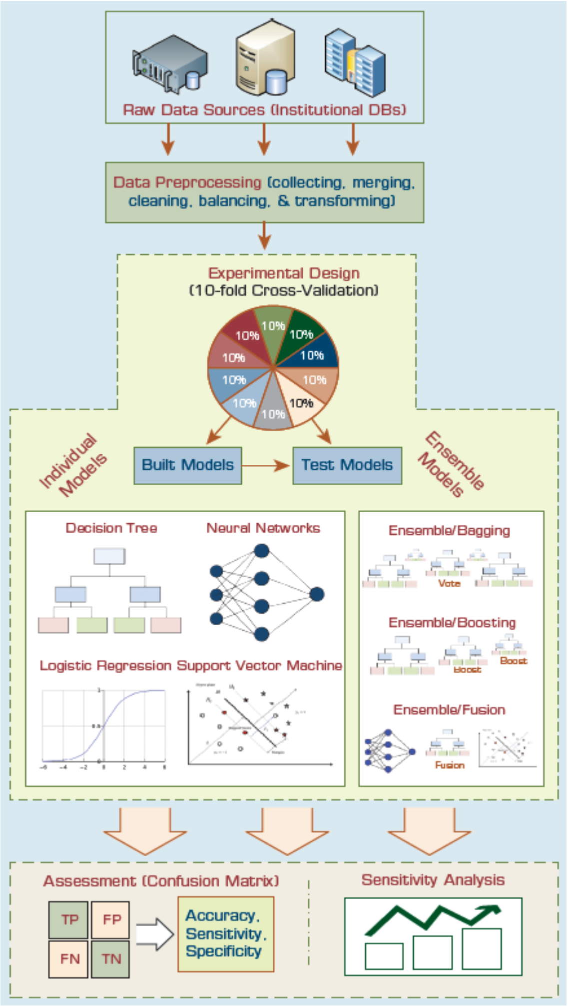
Improving Student Retention with Data-Driven Analytics

Questions for Discussion

1. What is student attrition, and why is it an important problem in higher education?
2. What were the traditional methods to deal with the attrition problem?
3. List and discuss the data-related challenges within context of this case study.
4. What was the proposed solution? And, what were the results?

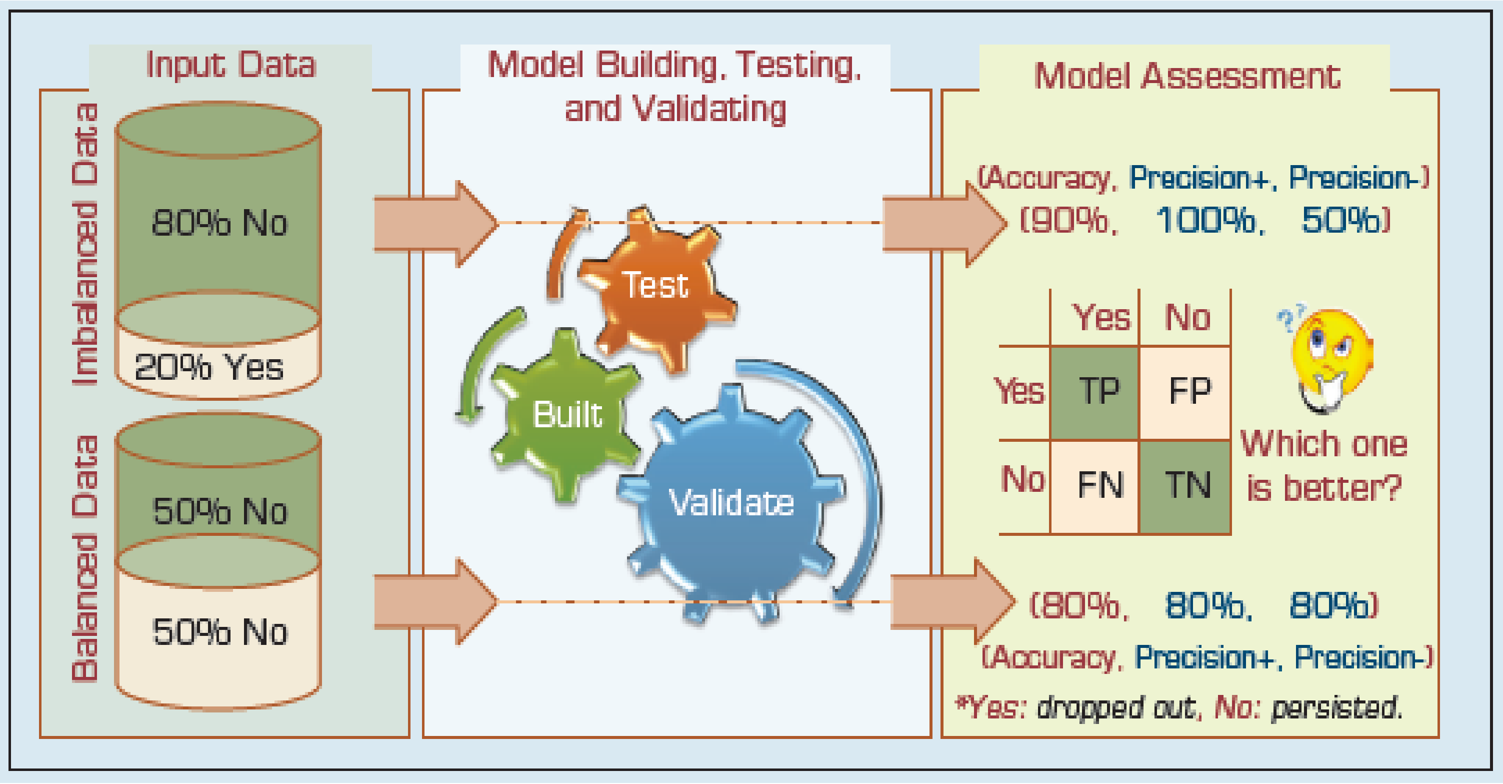
Application Case 2.2 (2 of 6)

- Student retention
  - Freshmen class
- Why it is important?
- What are the common techniques to deal with student attrition?
- Analytics versus theoretical approaches to student retention problem



Application Case 2.2 (3 of 6)

- Data imbalance problem



Application Case 2.2 (4 of 6)

Table 2.2 Prediction Results for the Original/Unbalanced Dataset

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1494	384	1518	304	1478	255	1438	376
Yes	1596	11142	1572	11222	1612	11271	1652	11150
SUM	3090	11526	3090	11526	3090	11526	3090	11526
Per-Class Accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall Accuracy	86.45%		87.16%		87.23%		86.12%	

Table 2.3 Prediction Results for the Balanced Data Set

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2309	464	2311	417	2313	386	2125	626
Yes	781	2626	779	2673	777	2704	965	2464
SUM	3090	3090	3090	3090	3090	3090	3090	3090
Per-class Accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall Accuracy	79.85%		79.85%		81.18%		74.26%	

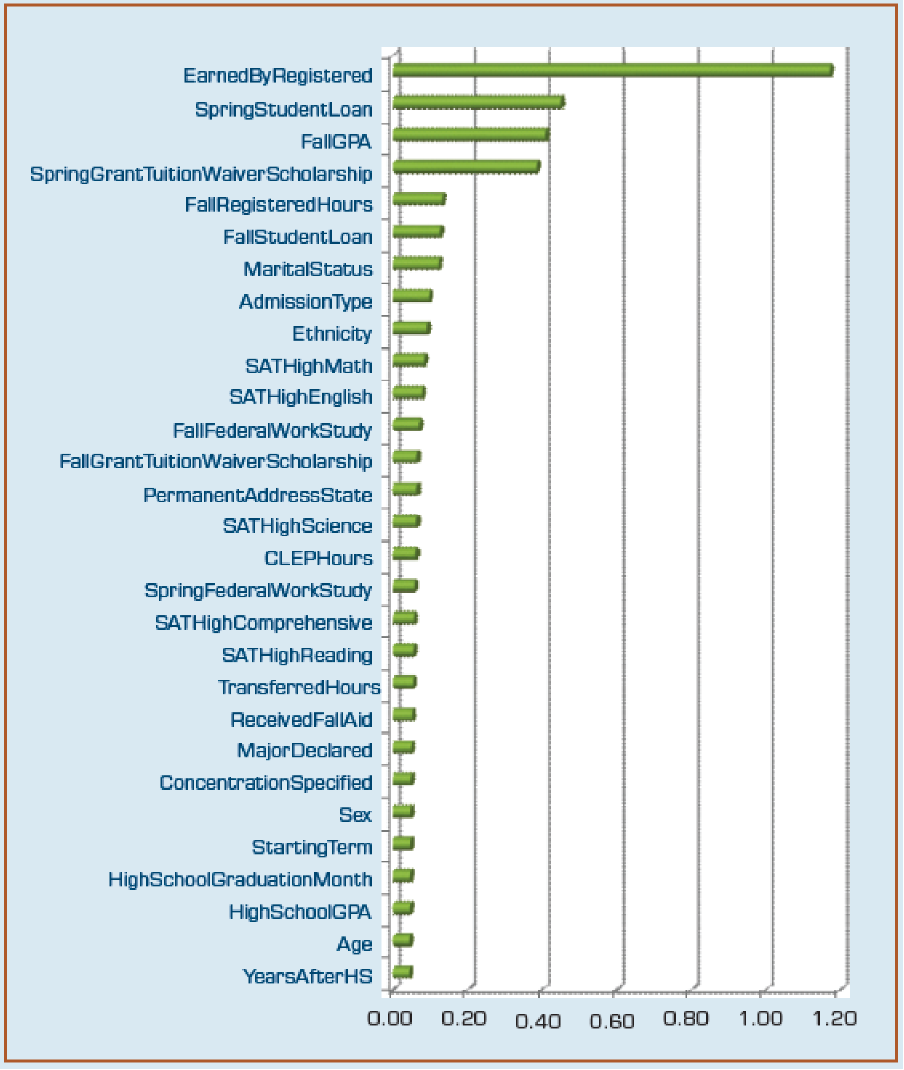
Application Case 2.2 (5 of 6)

Table 2.4 Prediction Results for the Three Ensemble Models

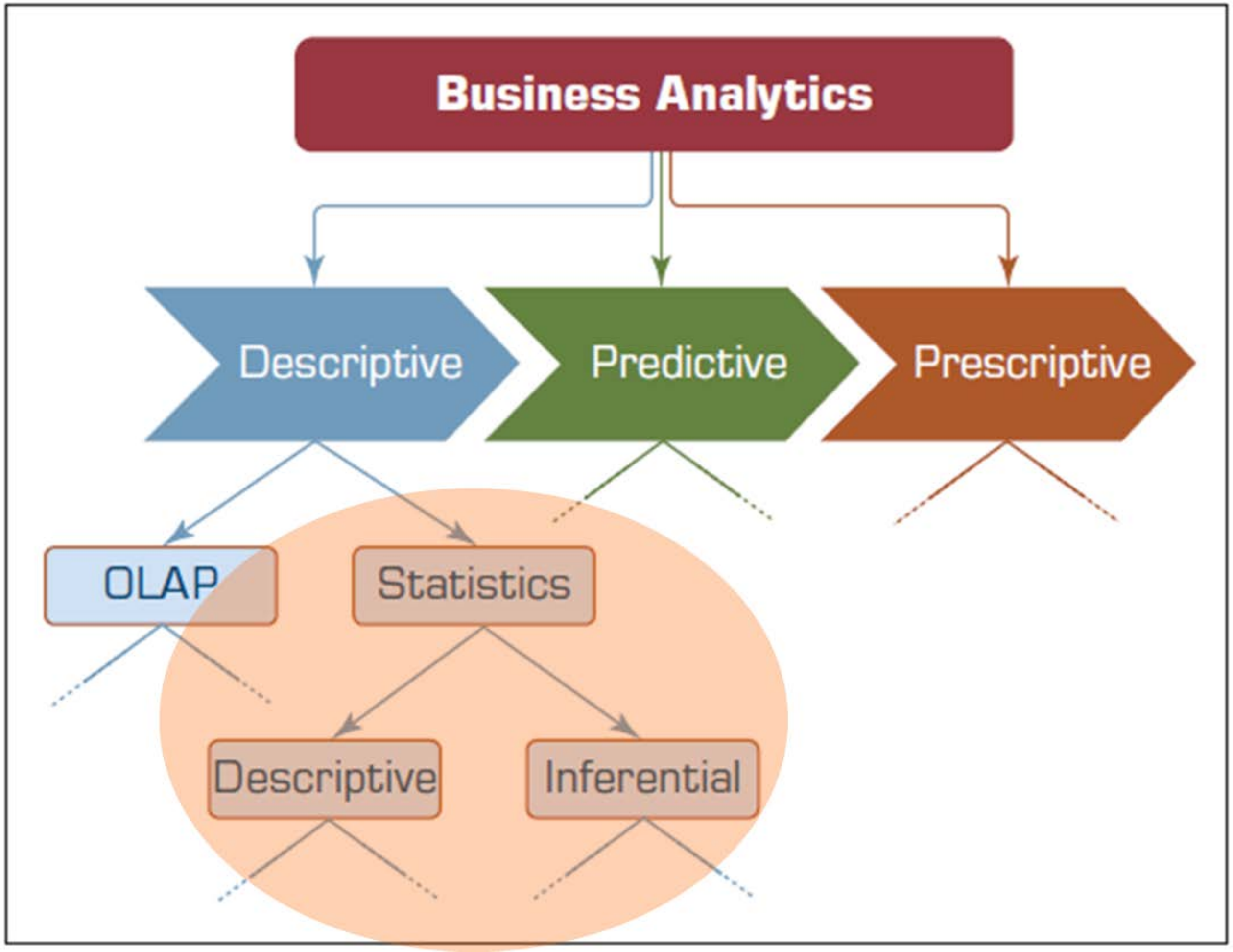
	Boosting		Bagging		Information Fusion	
	(Boosted Trees)		(Random Forest)		(Weighted Average)	
	No	Yes	No	Yes	No	Yes
No	2242	375	2327	362	2335	351
Yes	848	2715	763	2728	755	2739
SUM	3090	3090	3090	3090	3090	3090
Per-Class Accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall Accuracy	80.21%		81.80%		81.80%	

Application Case 2.2 (6 of 6)

- Results



Statistical Modeling for Business Analytics (1 of 2)



What is statistics?

Statistical Modeling for Business Analytics (2 of 2)

- **Statistics**
  - A collection of mathematical techniques to characterize and interpret data
- **Descriptive Statistics**
  - Describing the data (as it is)
- **Inferential statistics**
  - Drawing inferences about the population based on sample data
- Descriptive statistics for descriptive analytics

Descriptive Statistics Measures of Centrality Tendency

- **Arithmetic mean**

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Median**
  - The number in the middle
- **Mode**
  - The most frequent observation

Descriptive Statistics Measures of Dispersion (1 of 2)

- **Dispersion**
  - Degree of variation in a given variable
- **Range**
  - Max - Min
- **Variance**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})^2}{n-1}$$

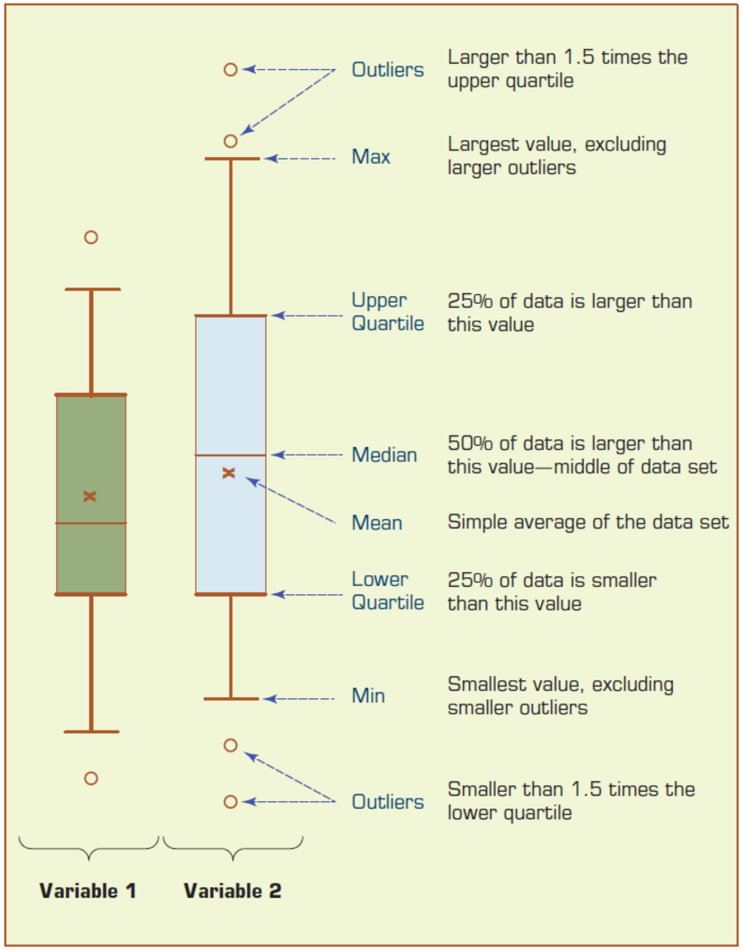
Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})^2}{n-1}}$$

- **Mean Absolute Deviation (MAD)**
  - Average absolute deviation from the mean

Descriptive Statistics Measures of Dispersion (2 of 2)

- Quartiles
- Box-and-Whiskers Plot
  – a.k.a. box-plot
  – Versatile / informative
  – Can show variation within data set



Descriptive Statistics Shape of a Distribution

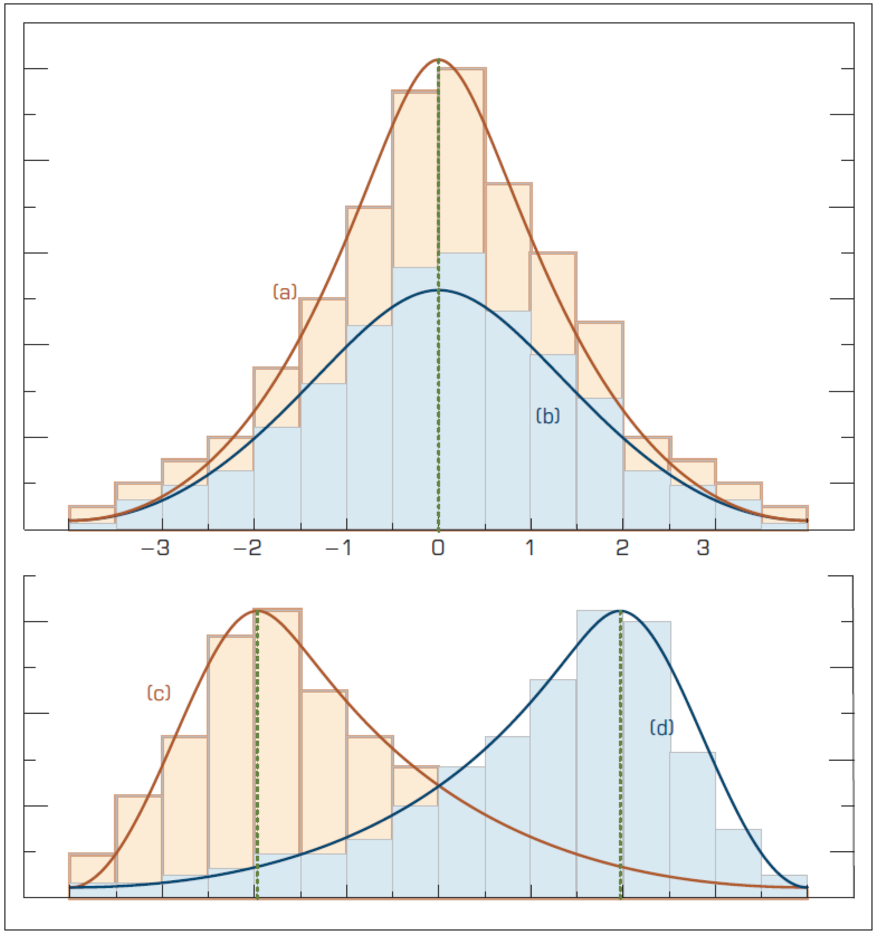
- **Histogram** – frequency chart
- **Skewness**
  - Measure of asymmetry

$$Skewness = S = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{(n-1)s^3}$$

- **Kurtosis**
  - Peak/tall/skinny nature of the distribution

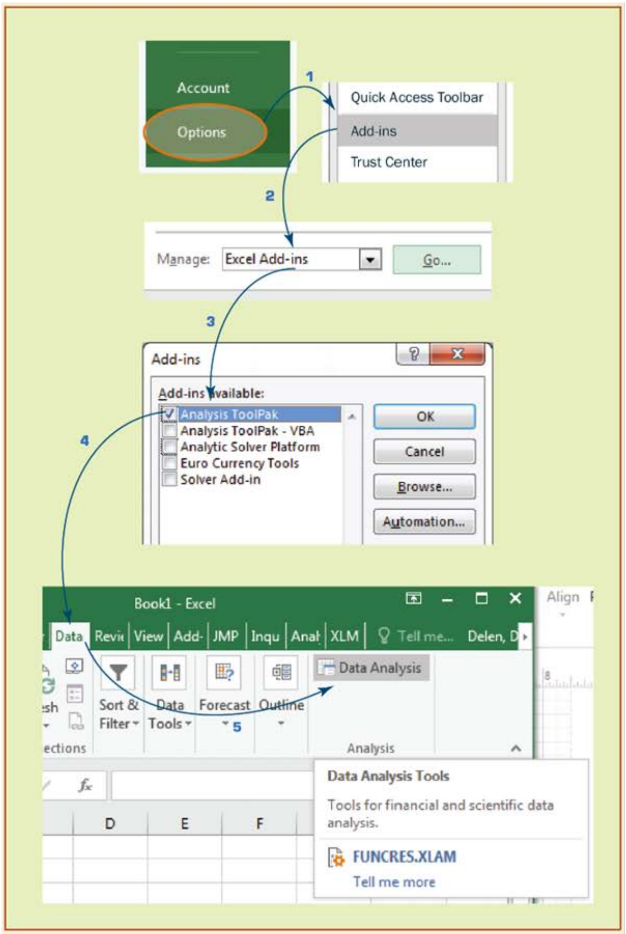
$$Kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{ns^4} - 3$$

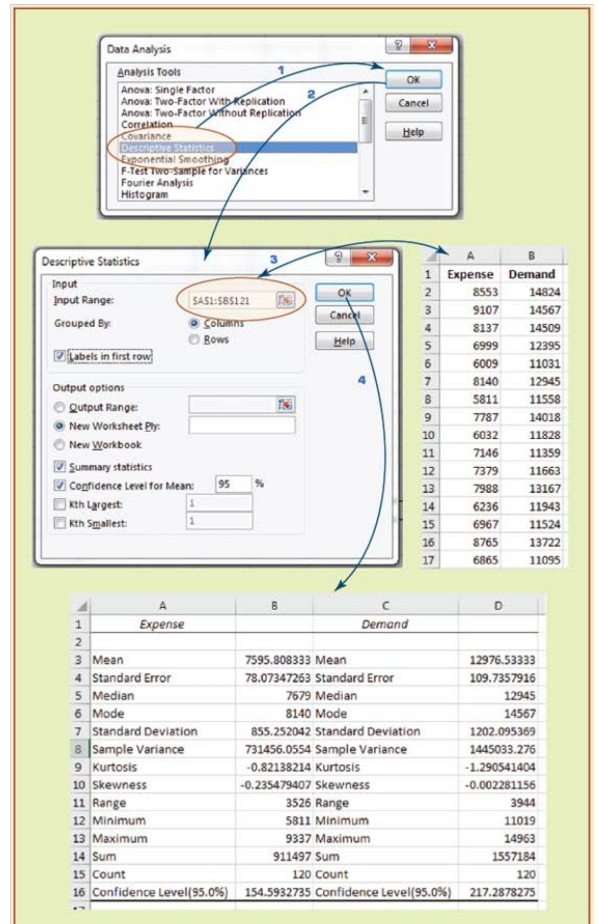
Relationship Between Dispersion and Shape Properties



Technology Insights 2.1 (1 of 2)

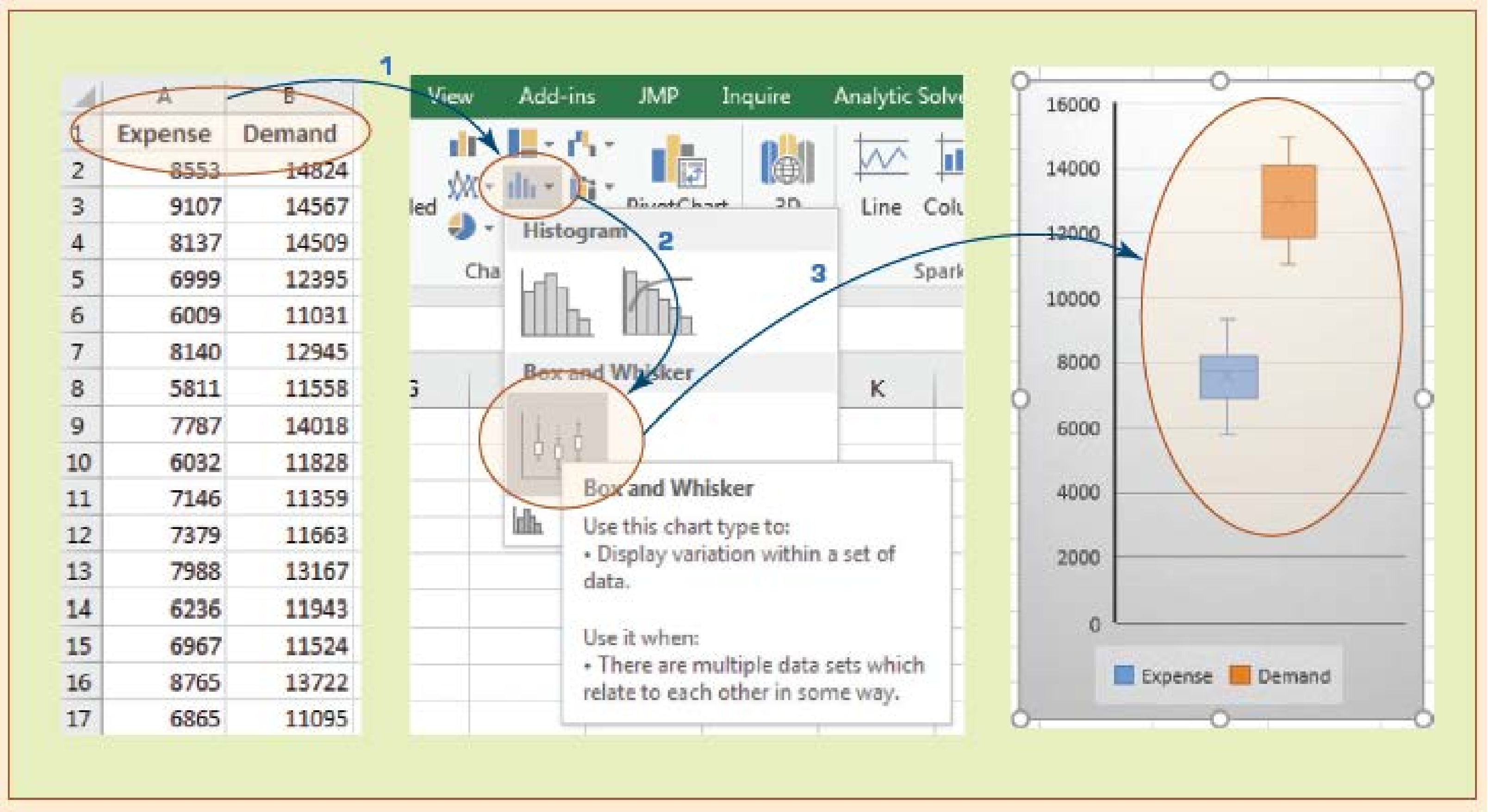
Descriptive Statistics in Excel





Technology Insights 2.1 (2 of 2)

Descriptive Statistics in Excel Creating box-plot in Microsoft Excel



Application Case 2.3

Town of Cary Uses Analytics to Analyze Data from Sensors, Assess Demand, and Detect Problems

Questions for Discussion

1. What were the challenges the Town of Cary was facing?
2. What was the proposed solution?
3. What were the results?
4. What other problems and data analytics solutions do you foresee for towns like Cary?

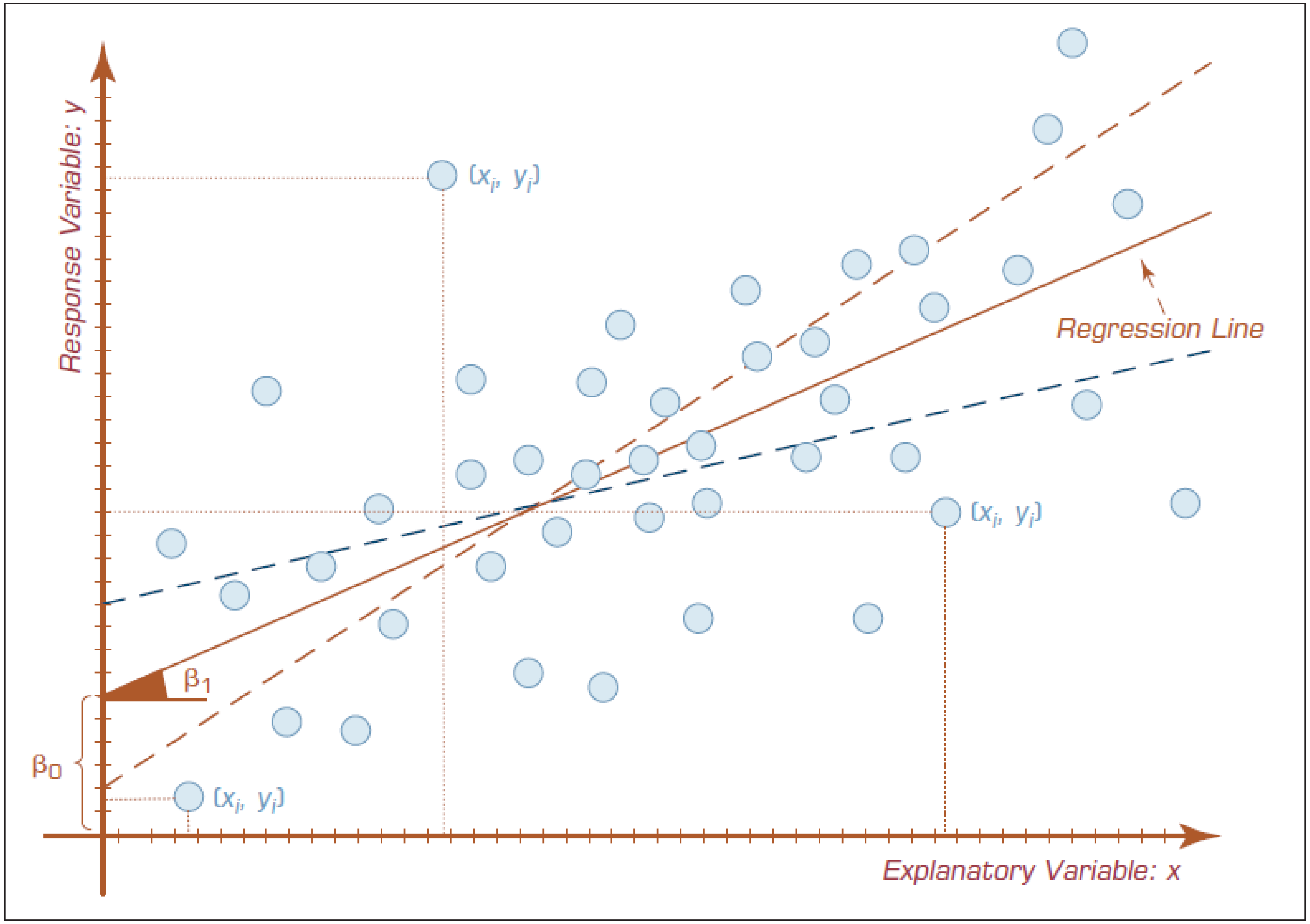
Regression Modeling for Inferential Statistics

- **Regression**
  - A part of inferential statistics
  - The most widely known and used analytics technique in statistics
  - Used to characterize relationship between explanatory (input) and response (output) variable
- It can be used for
  - Hypothesis testing (explanation)
  - Forecasting (prediction)

Regression Modeling (1 of 3)

- Correlation versus Regression
  - What is the difference (or relationship)?
- Simple Regression versus Multiple Regression
  - Base on number of input variables
- How do we develop linear regression models?
  - Scatter plots (visualization—for simple regression)
  - Ordinary least squares method
    - A line that minimizes squared of the errors

Regression Modeling (2 of 3)



Regression Modeling (3 of 3)

- *x*: input, *y*: output
- Simple Linear Regression

$$y = \beta_0 + \beta_1 x$$

- Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

- The meaning of Beta (β) coefficients
  - Sign (+ or -) and magnitude

Example: Linear Regression 1 of 2

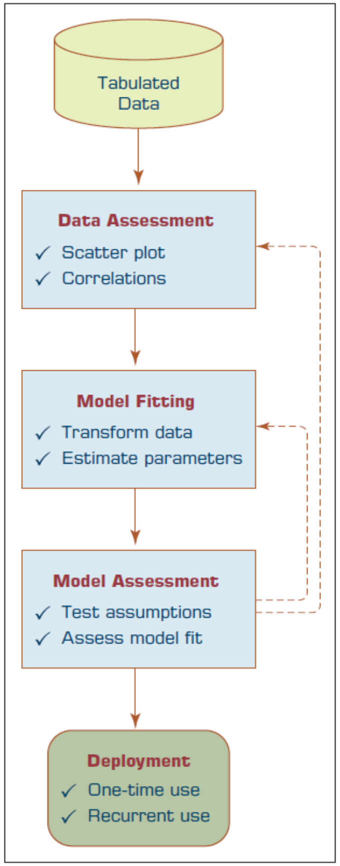
Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

- Want to predict Sales. If we use advertising as the predictor variable, linear regression estimates that
- **Sales = 168 + 23 Advertising**.
- That is, if advertising expenditure is increased by one Euro, then sales will be expected to increase by 23 million Euro, and if there was no advertising we would expect sales of 168 million Euro.

Process of Developing a Regression Model

How do we know if the model is good enough?

- **R^2** (R-Square)
- ***p*** Values
- Error measures (for prediction problems)
  - MSE, MAD, RMSE



Example: Linear Regression

- R-square = 0.98

Linear Regression: Sales

	Estimate	Standard Error	*t*	*p*
(Intercept)	167.68	58.94	2.85	.025
Advertising	23.42	1.37	17.13	< .001

n = 9 cases used in estimation; R-squared: 0.9767; Correct predictions: 88.89%; AIC: 100.34; multiple comparisons correction: None

Regression Modeling Assumptions

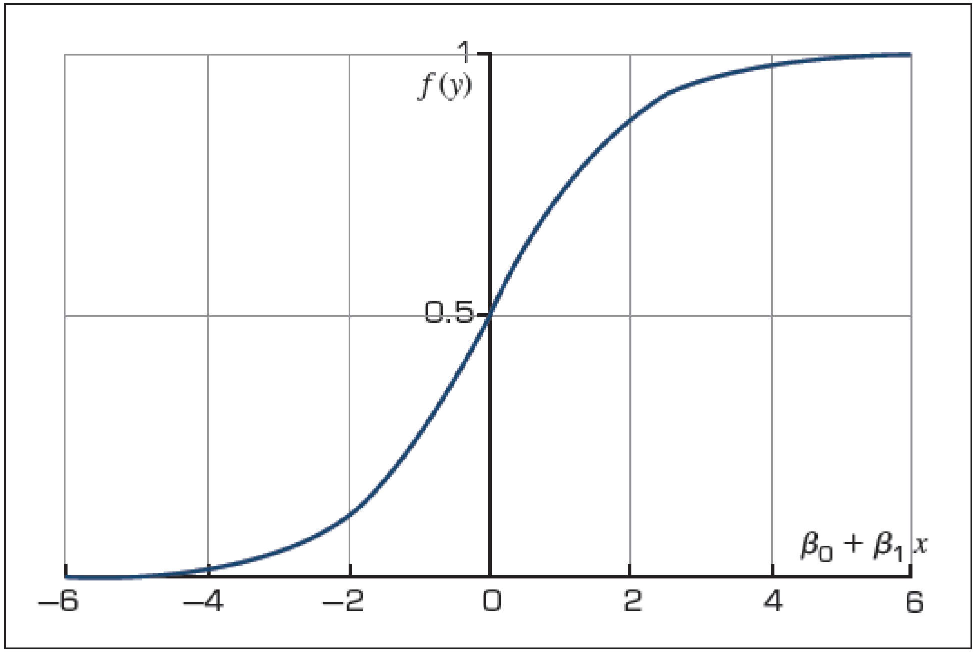
- Linearity (Linear regression)
- Independence
- Normality (Normal Distribution)
- Constant Variance
- Multicollinearity
- What happens if the assumptions do Not hold?
  - What do we do then?

Logistic Regression Modeling (1 of 2)

- A very popular statistics-based classification algorithm
- Employs supervised learning
- Developed in 1940s
- The difference between Linear Regression and Logistic Regression
  - In Logistic Regression Output/Target variable is a binomial (binary classification) variable (as opposed to numeric variable)

Logistic Regression Modeling (2 of 2)

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

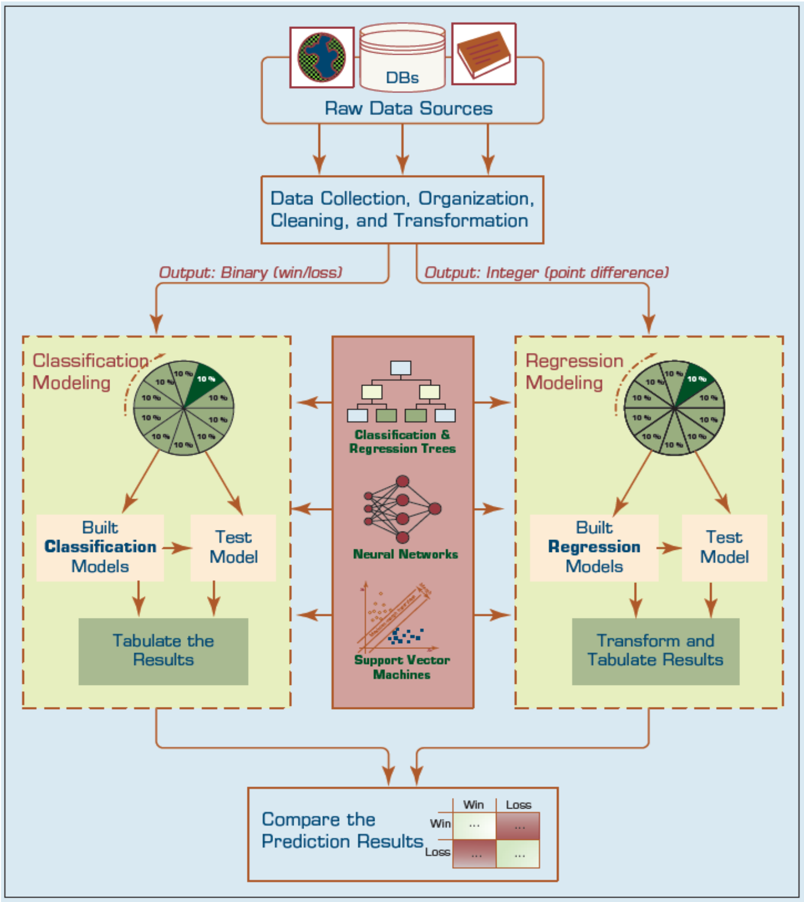


Application Case 2.4 (1 of 6)

Predicting NCAA Bowl Game Outcomes

Application Case 2.4 (2 of 6)

- The analytics process to develop prediction models (both regression and classification type) for NCAA Bowl Game outcomes



Application Case 2.4 (3 of 6)

Prediction Results

1. Classification
2. Regression

Application Case 2.4 (4 of 6)

Table 2.6 Prediction results for the direct classification methodology

Prediction Method (Classification*)		Confusion Matrix		Accuracy** (in %)	Sensitivity (in %)	Specificity (in %)
		Win	Loss			
ANN (MLP)	Win	92	42	75.00	68.66	82.73
	Loss	19	91			
SVM (RBF)	Win	105	29	79.51	78.36	78.36
	Loss	21	89			
DT (C&RT)	Win	113	21	**86.48**	84.33	89.09
	Loss	12	98			

*The output variable is **a binary categorical variable (Win or Loss)**; differences were sig (** $p < 0.01$).

Application Case 2.4 (5 of 6)

Table 2.7 Prediction results for the regression-based classification methodology

Prediction Method (Regression-Based*)		Confusion Matrix		Accuracy**	Sensitivity	Specificity
		Win	Loss			
ANN (MLP)	Win	94	40	72.54	70.15	75.45
	Loss	27	83			
SVM (RBF)	Win	100	34	74.59	74.63	74.55
	Loss	28	82			
DT (C&RT)	Win	106	28	77.87	76.36	79.10
	Loss	26	84			

*The output variable is **a numerical/integer variable (point-diff)**; differences were sig (** $p < 0.01$).

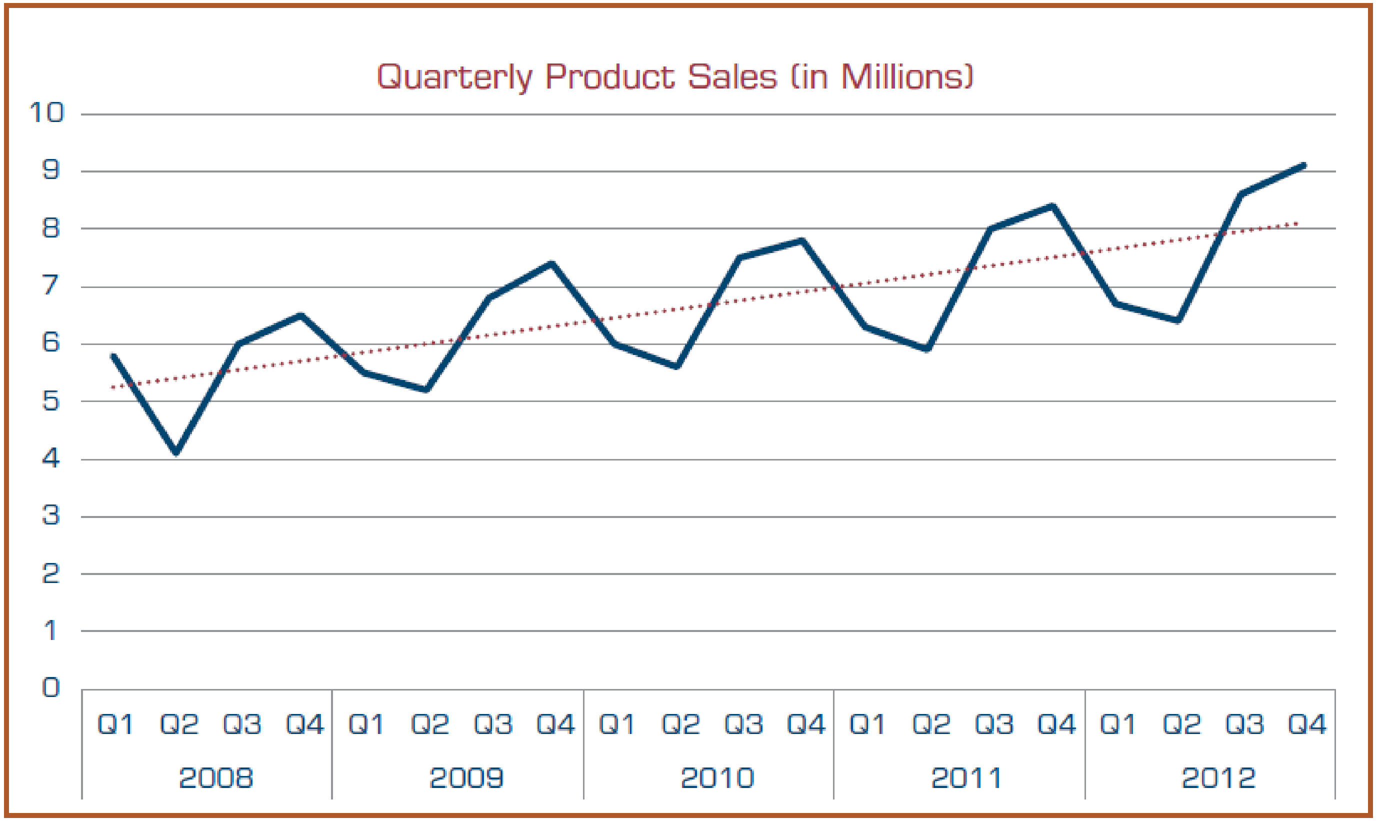
Application Case 2.4 (6 of 6)

Questions for Discussion

1. What are the foreseeable challenges in predicting sporting event outcomes (e.g., college bowl games)?
2. How did the researchers formulate/design the prediction problem (i.e., what were the inputs and output, and what was the representation of a single sample—row of data)?
3. How successful were the prediction results? What else can they do to improve the accuracy?

Time Series Forecasting

- Is it different than Simple Linear Regression? How?



Business Reporting Definitions and Concepts

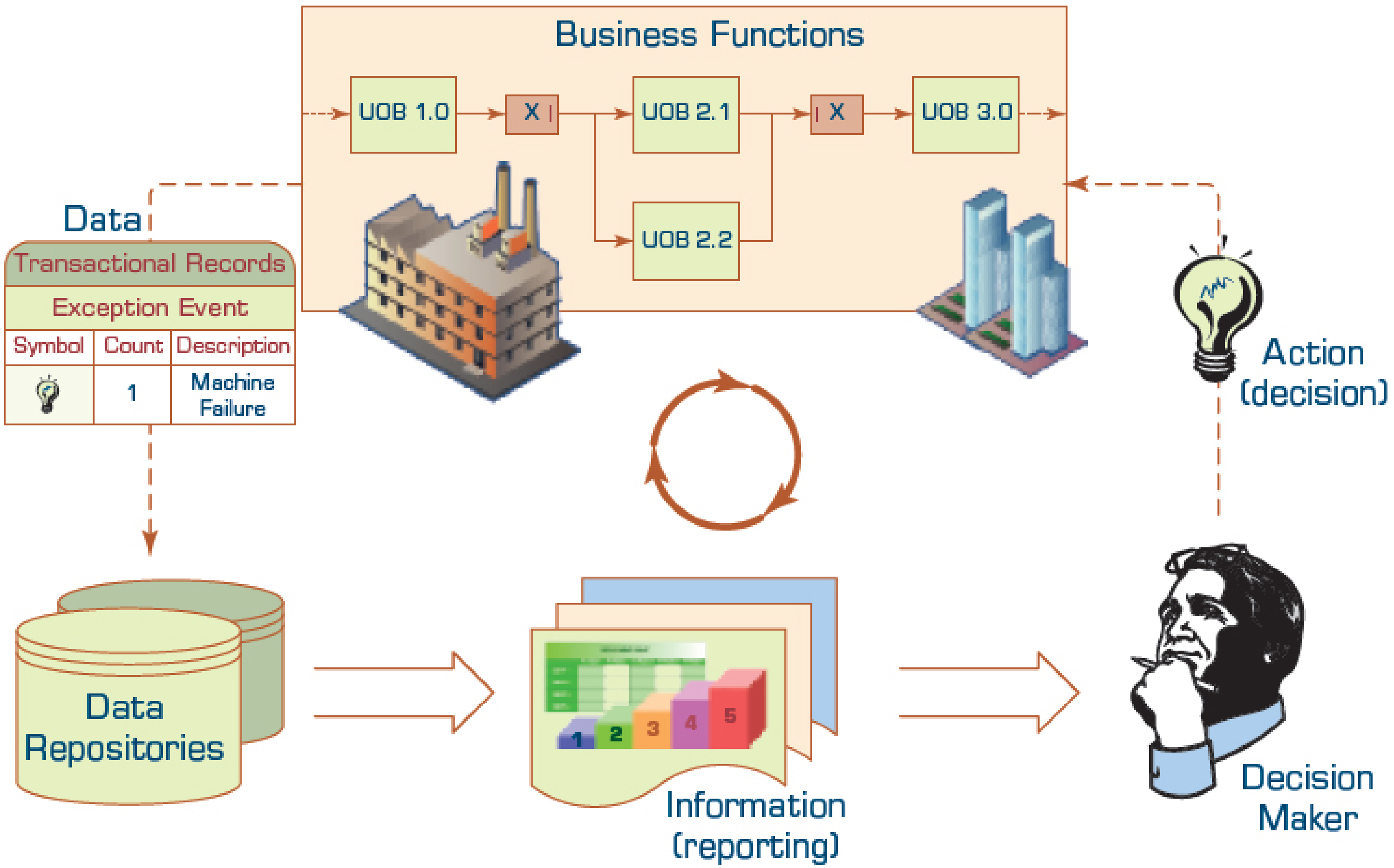
- Report = Information → Decision
- Report?
  - Any communication artifact prepared to convey specific information
- A report can fulfill many functions
  - To ensure proper departmental functioning
  - To provide information
  - To provide the results of an analysis
  - To persuade others to act
  - To create an organizational memory…

What is a Business Report?

- A written document that contains information regarding business matters.
- **Purpose:** to improve managerial decisions
- **Source:** data from inside and outside the organization (via the use of ETL)
- **Format:** text + tables + graphs/charts
- **Distribution:** in-print, email, portal/intranet

Data acquisition → Information generation → Decision making → Process management

Business Reporting



Types of Business Reports

- Metric Management Reports
  - Help manage business performance through metrics (SLAs for externals; KPIs for internals)
  - Can be used as part of Six Sigma and/or TQM
- Dashboard-Type Reports
  - Graphical presentation of several performance indicators in a single page using dials/gauges
- Balanced Scorecard–Type Reports
  - Include financial, customer, business process, and learning & growth indicators

Application Case 2.5

Flood of Paper Ends at FEMA (Federal Emergency Management Agency)

Questions for Discussion

1. What does FEMA do?

   - help people before, during and after disasters.

2. What are the main challenges that FEMA faces?
3. How did FEMA improve its inefficient reporting practices?

   -WebFOCUS solution

What is data visualization?

Data Visualization VS Information Visualization?

Data Visualization

“The use of visual representations to explore, make sense of, and communicate data.”

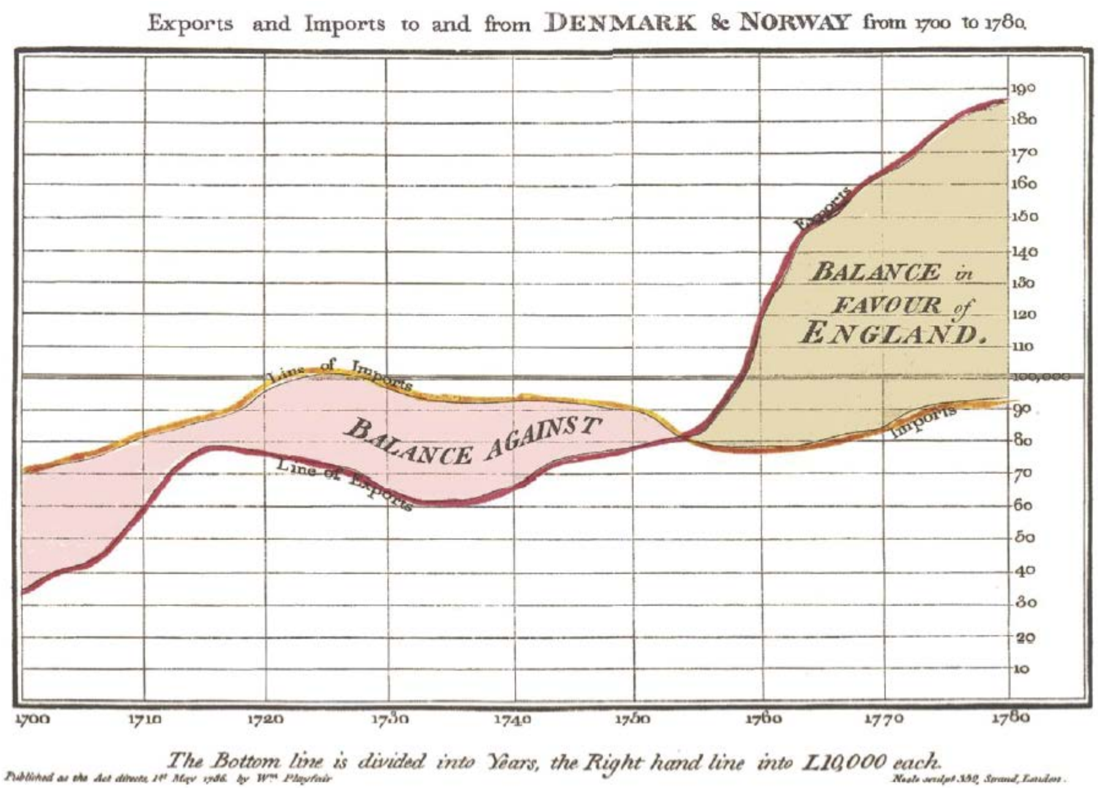
- Data visualization vs. Information visualization
- Information = aggregation, summarization, and contextualization of data
- Related to information graphics, scientific visualization, and statistical graphics
- Often includes charts, graphs, illustrations, …

A Brief History of Data Visualization

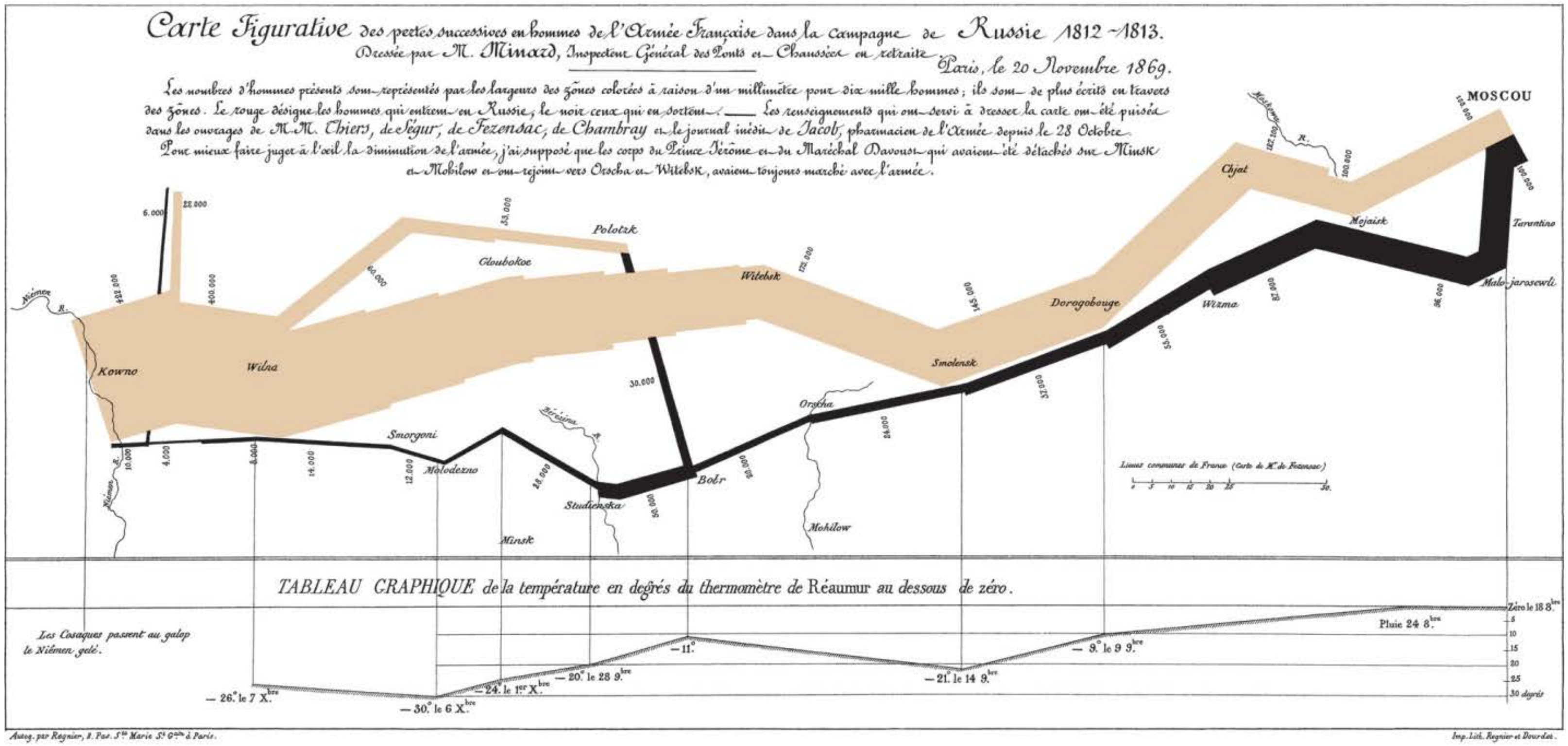
- Data visualization can date back to the second century AD
- Most developments have occurred in the last two and a half centuries
- Until recently it was not recognized as a discipline
- Today’s most popular visual forms date back a few centuries

The First Pie Chart Created by William Playfair in 1801

William Playfair is widely credited as the inventor of the modern chart, having created the first line and pie charts.



Decimation of Napoleon’s Army During the 1812 Russian Campaign



By Charles Joseph Minard

- Arguably the most popular multi-dimensional chart

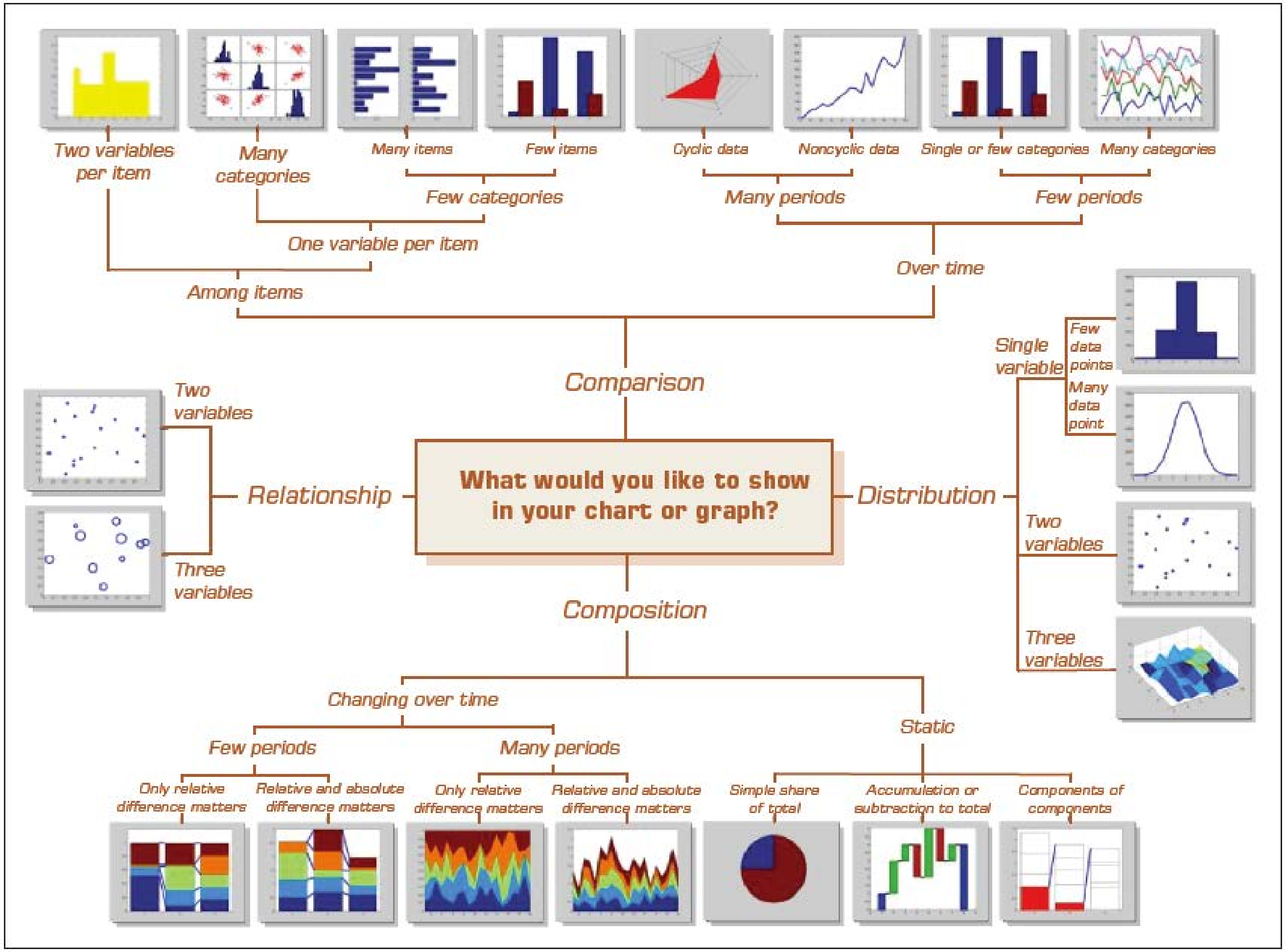
Application Case 2.6

Macfarlan Smith Improves Operational Performance Insight with Tableau Online

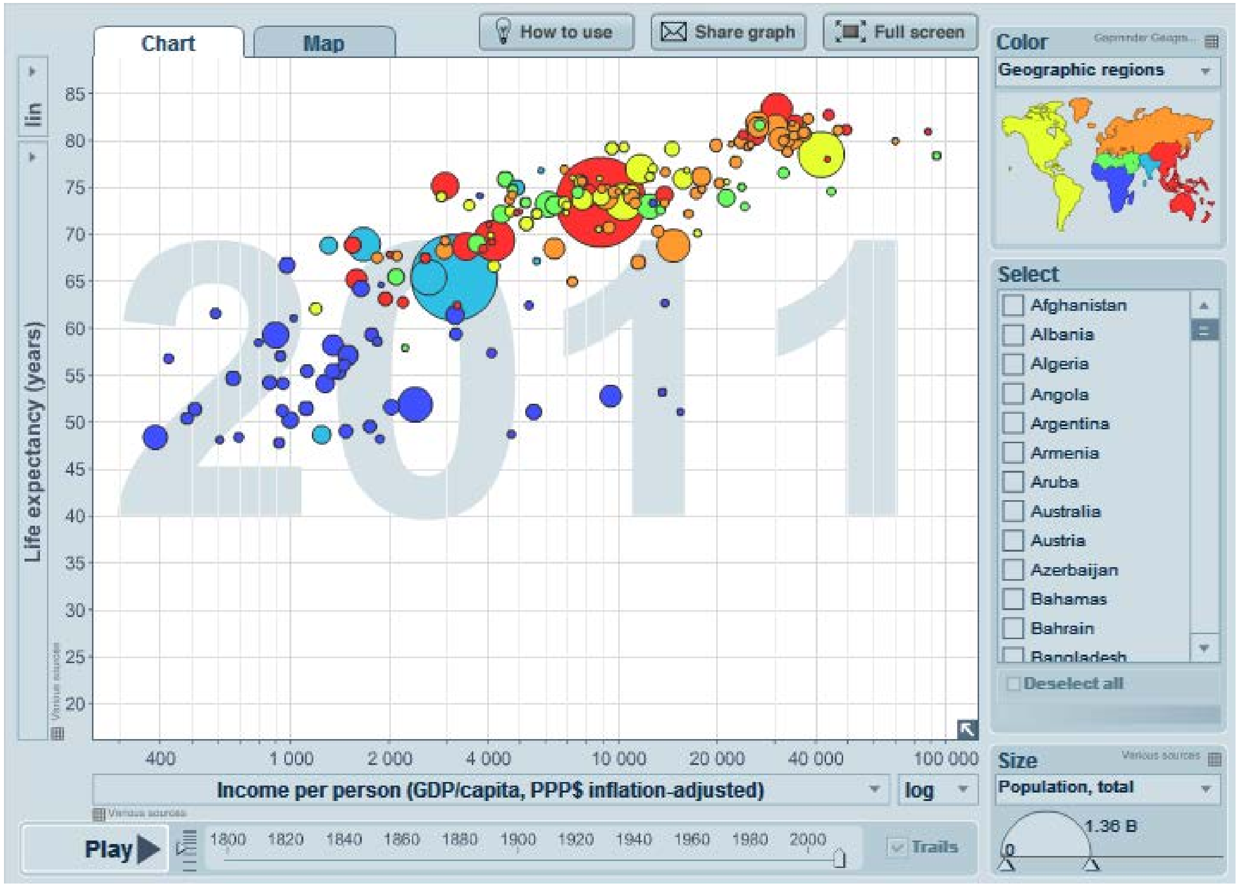
Questions for Discussion

1. What were the data and reporting related challenges Macfarlan Smith facing?
2. What was the solution and the obtained results and/or benefits?

Which Chart or Graph Should You Use?



An Example Gapminder Chart Wealth and Health of Nations



See gapminder.org for Interesting animated examples

The Emergence of Data Visualization and Visual Analytics (1 of 2)

- Magic Quadrant for Business Intelligence and Analytics Platforms (Source: Gartner.com)
- Many data visualization companies are in the 4th quadrant
- There is a move towards visualization

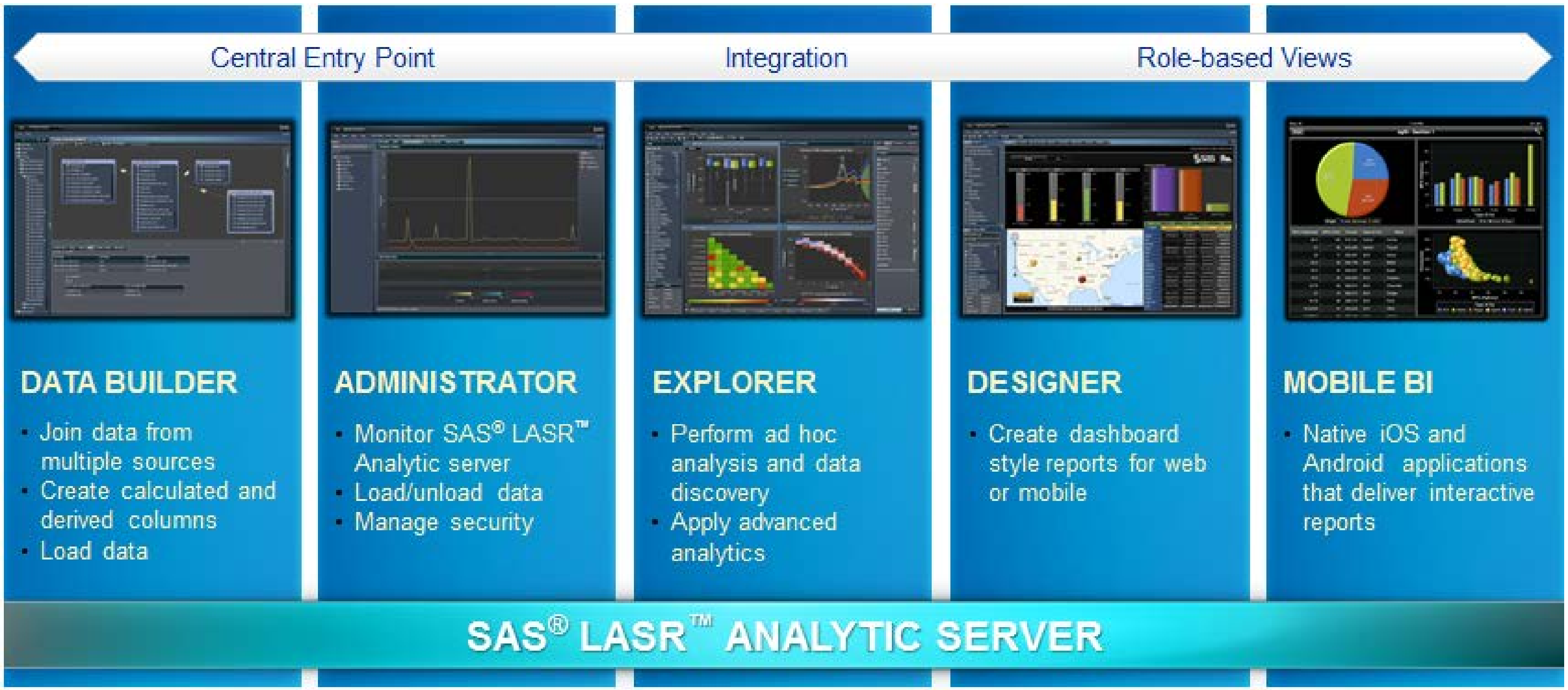
The Emergence of Data Visualization and Visual Analytics (2 of 2)

- Emergence of new companies
  - Tableau, Spotfire, QlikView, …
- Increased focus by the big players
  - MicroStrategy improved Visual Insight
  - SAP launched Visual Intelligence
  - SAS launched Visual Analytics
  - Microsoft bolstered PowerPivot with Power View
  - IBM launched Cognos Insight
  - Oracle acquired Endeca

Visual Analytics

- A recently coined term
  - Information visualization + predictive analytics
- Information visualization
  - Descriptive, backward focused
  - “what happened” “what is happening”
- Predictive analytics
  - Predictive, future focused
  - “what will happen” “why will it happen”
- There is a strong move toward **visual analytics**

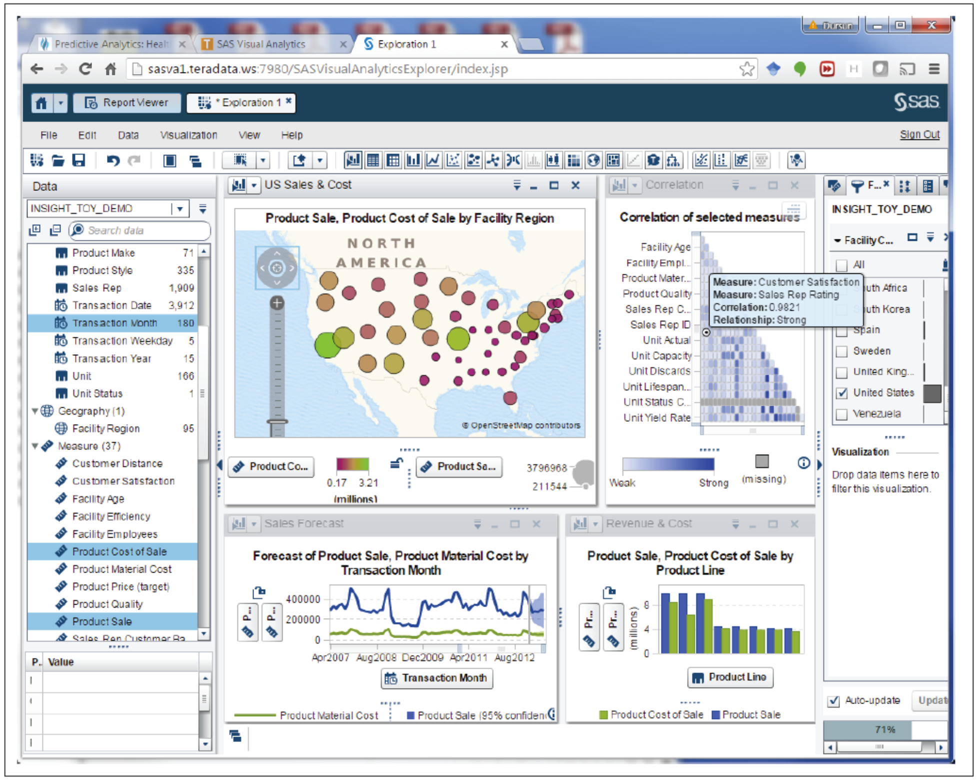
Visual Analytics by SAS Institute (1 of 2)



- SAS Visual Analytics Architecture
  – Big data + In memory + Massively parallel processing + ..

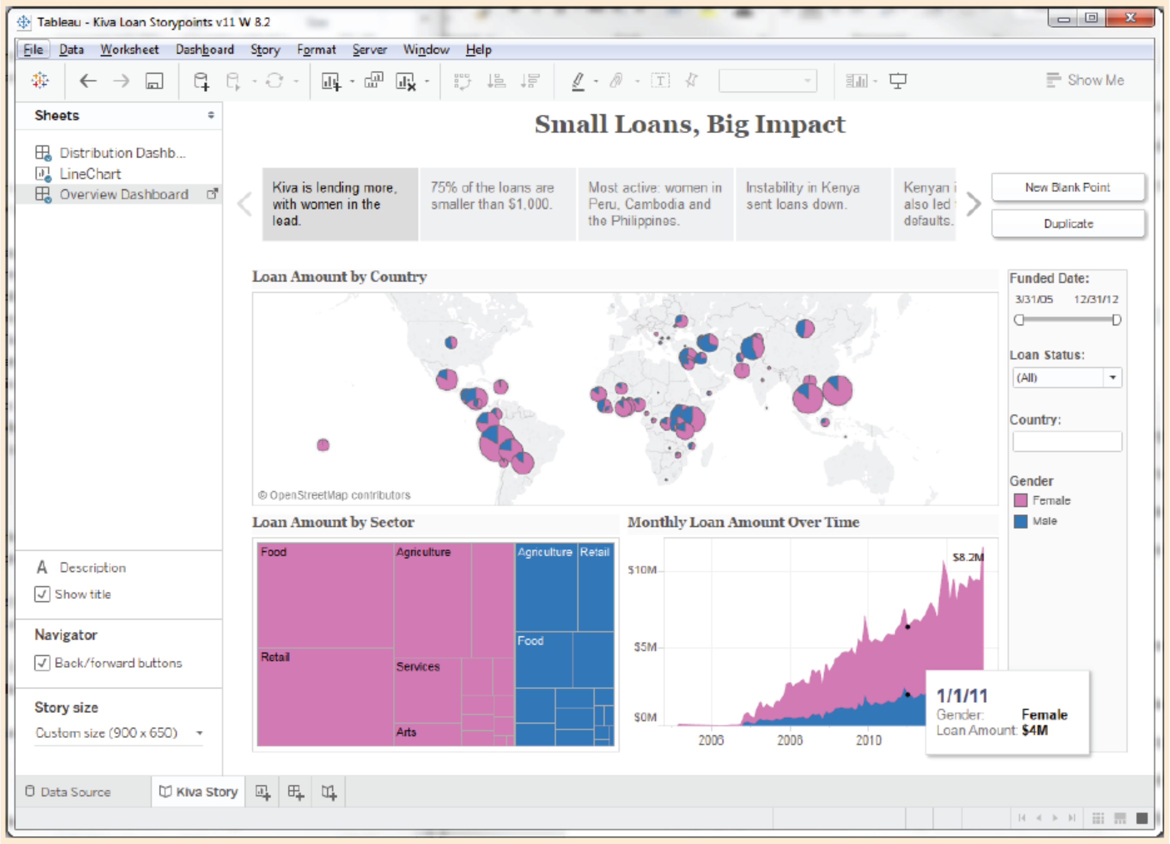
Visual Analytics by SAS Institute (2 of 2)

- At teradatauniversitynetwork.com, you can learn more about SAS VA, experiment with the tool



Technology Insight 2.3

Telling Great Stories with Data and Visualization

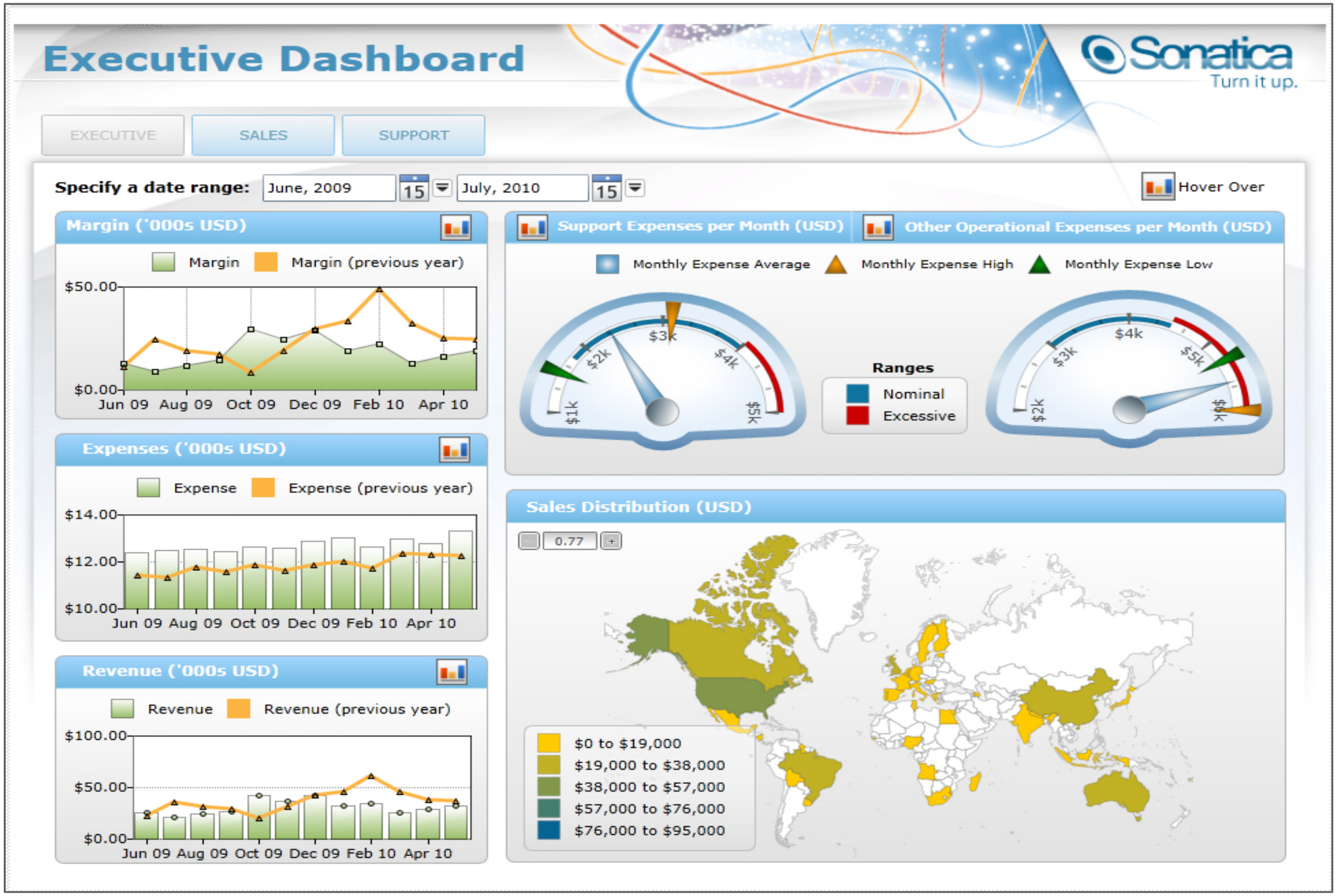


What is performance dashboard?

Performance Dashboards (1 of 4)

- Performance dashboards are commonly used in BPM software suites and BI platforms
- Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored

Performance Dashboards (2 of 4)



Application Case 2.7

Dallas Cowboys Score Big with Tableau and Teknion

Questions for Discussion

1. How did the Dallas Cowboys use information visualization?
2. What were the challenge, the proposed solution, and the obtained results?

Performance Dashboards (3 of 4)

- Dashboard design
  – The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly
- Three layer of information
  – Monitoring
  – Analysis
  – Management

Performance Dashboards (4 of 4)

- What to look for in a dashboard
  - Use of visual components to highlight data and exceptions that require action
  - Transparent to the user, meaning that they require minimal training and are extremely easy to use
  - Combine data from a variety of systems into a single, summarized, unified view of the business
  - Enable drill-down or drill-through to underlying data sources or reports
  - Present a dynamic, real-world view with timely data
  - Require little coding to implement, deploy, and maintain

Best Practices in Dashboard Design

- Benchmark KPIs with Industry Standards
- Wrap the Metrics with Contextual Metadata
- Validate the Design by a Usability Specialist
- Prioritize and Rank Alerts and Exceptions
- Enrich Dashboard with Business-User Comments
- Present Information in Three Different Levels
- Pick the Right Visual Constructs
- Provide for Guided Analytics

Application Case 2.8

Visual Analytics Helps Energy Supplier Make Better Connections

Questions for Discussion

1. Why do you think energy supply companies are among the prime users of information visualization tools?
2. How did Electrabel use information visualization for the single version of the truth?
3. What were their challenges, the proposed solution, and the obtained results?