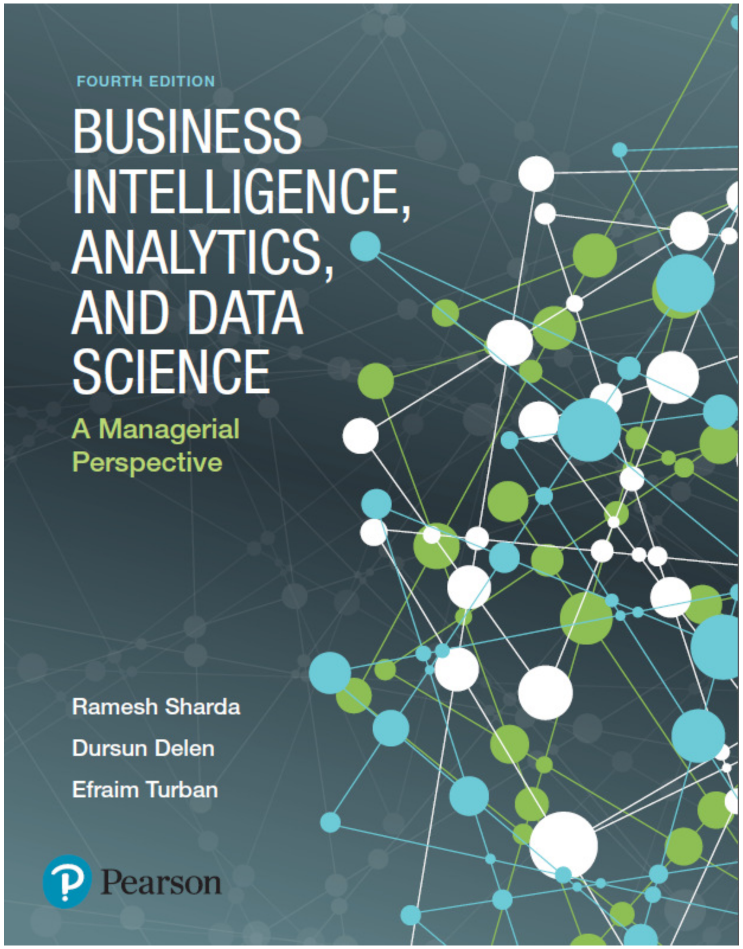
Business Intelligence, Analytics, and Data Science: A Managerial Perspective

Fourth Edition



Chapter 2 – Part A

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

OPENING VIGNETTE

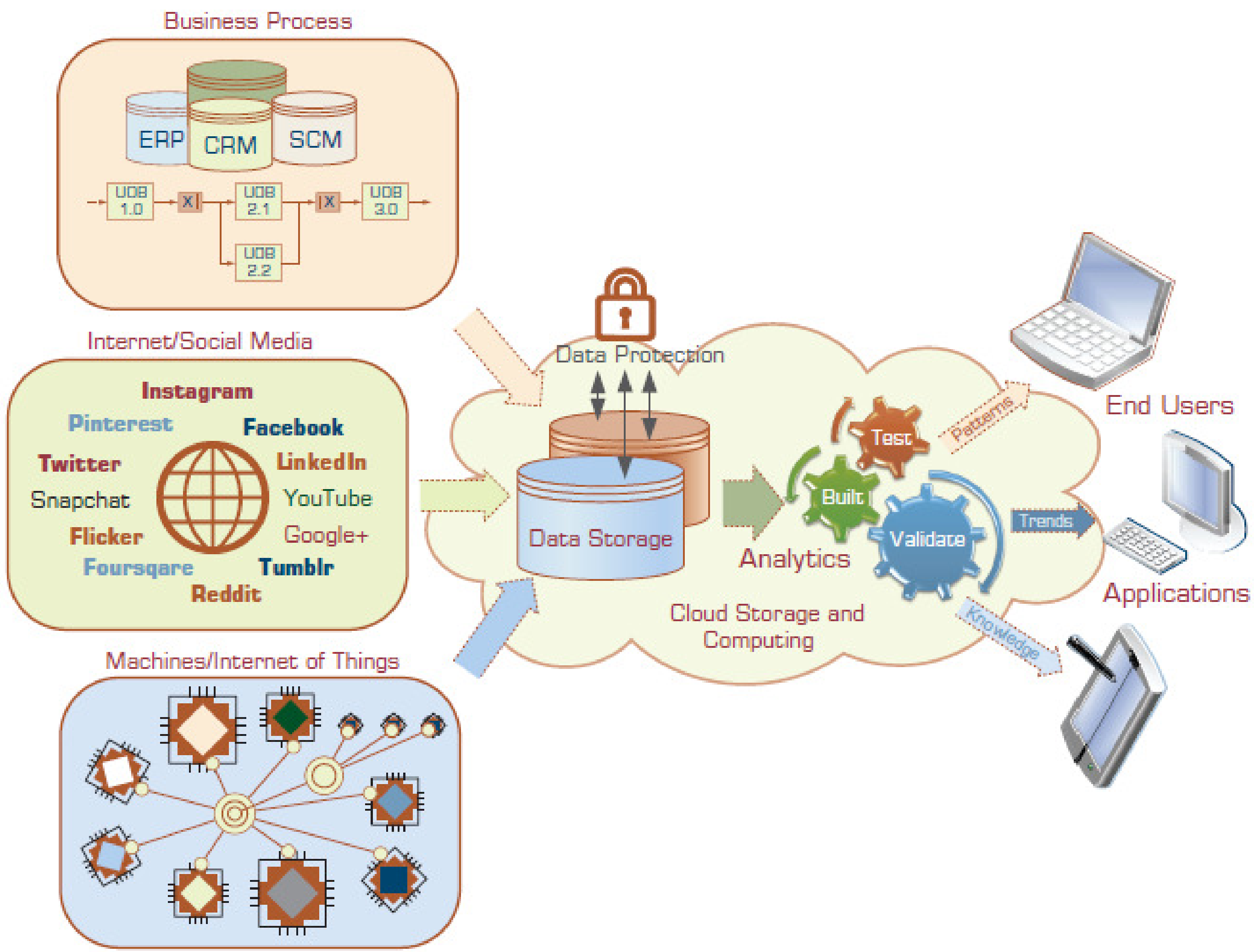
Attracts and Engages a New Generation of Radio Consumers with Data-Driven Marketing

1. What does SiriusXM do? In what type of market does it conduct its business?
2. What were the challenges? Comment on both technology and data-related challenges.
3. What were the proposed solutions?
4. How did they implement the proposed solutions? Did they face any implementation challenges?
5. What were the results and benefits? Were they worth the effort/investment?

The Nature of Data

- 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



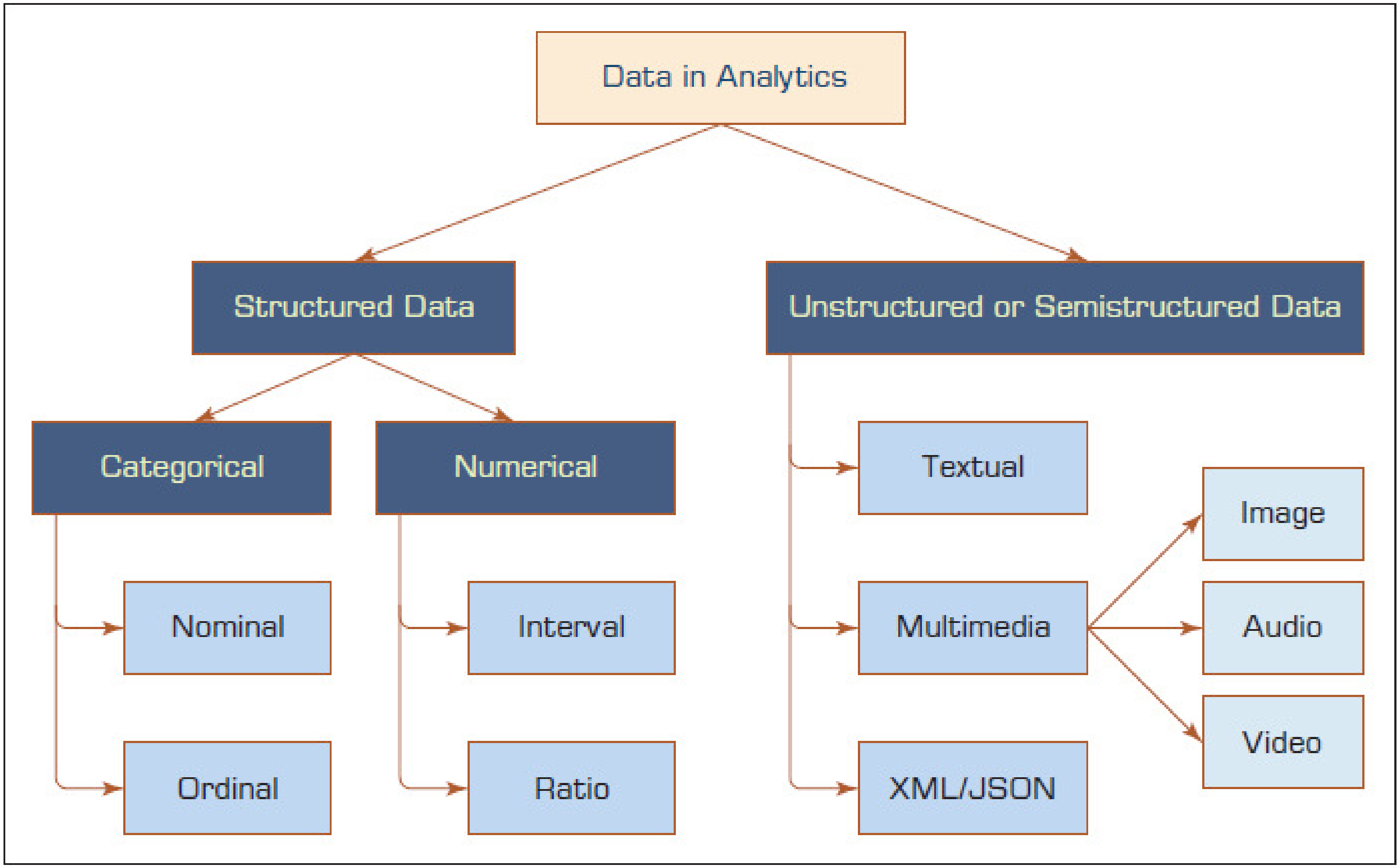
Metrics for Analytics Ready Data

- 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

A Simple Taxonomy of Data

- 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



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)?

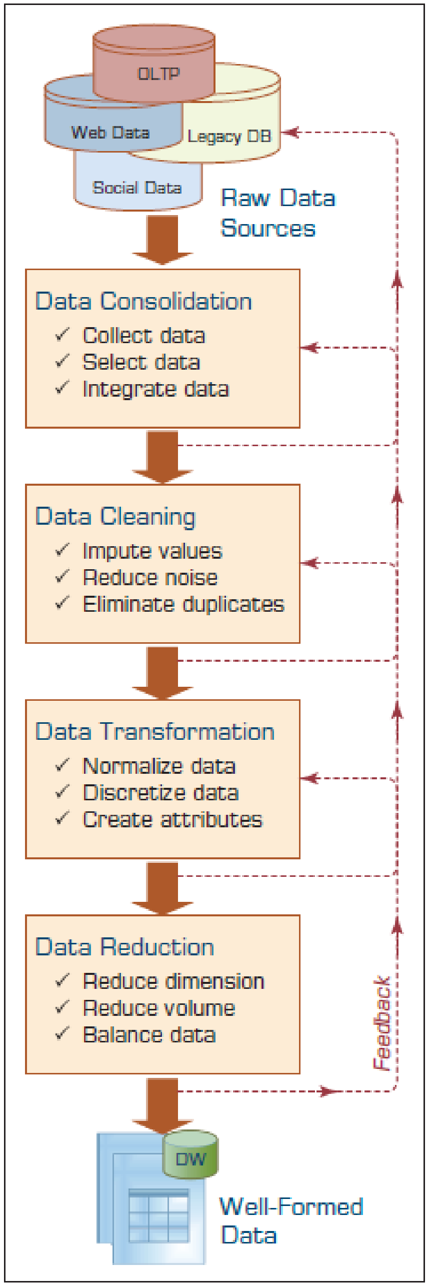
The Art and Science of Data Preprocessing

- 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

- Data reduction

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



Data Preprocessing Tasks and Methods

TABLE 2.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	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.
	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.
	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 4)

Improving Student Retention with Data-Driven Analytics

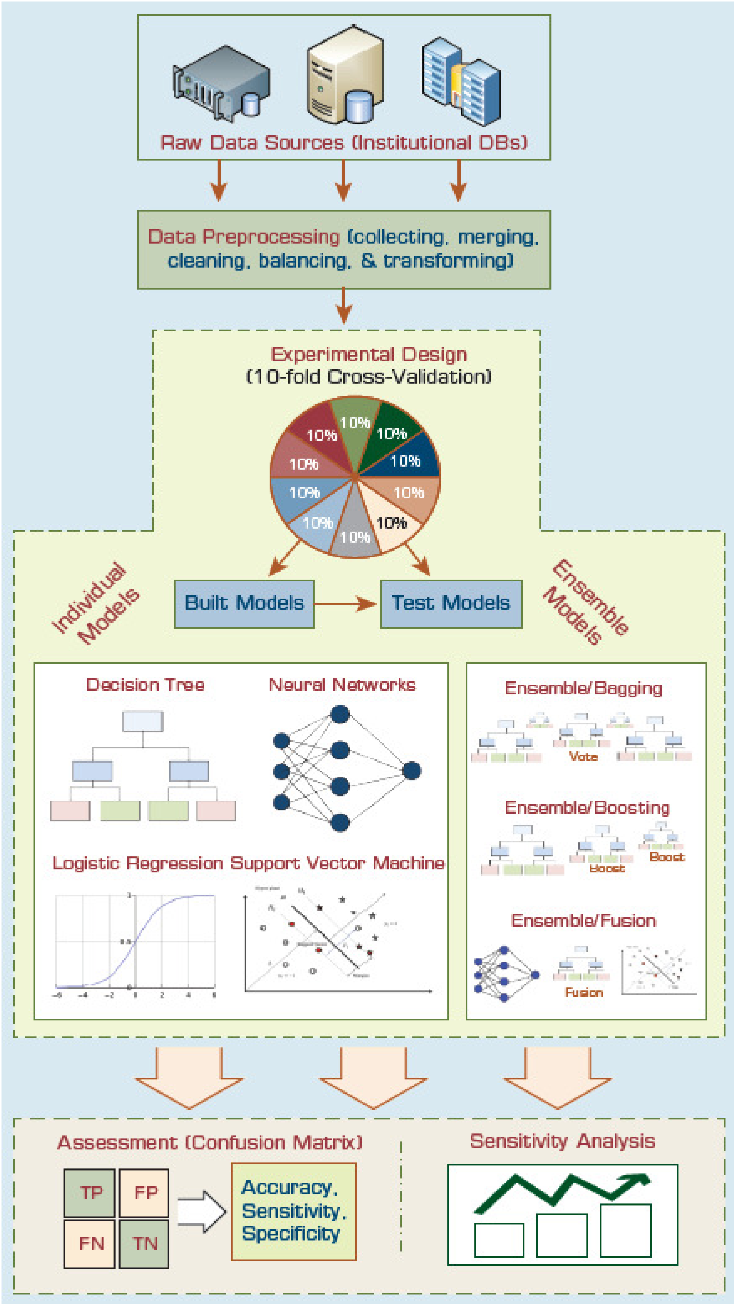
Questions for Discussion

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

Improving Student Retention with Data-Driven Analytics (2 of 4)

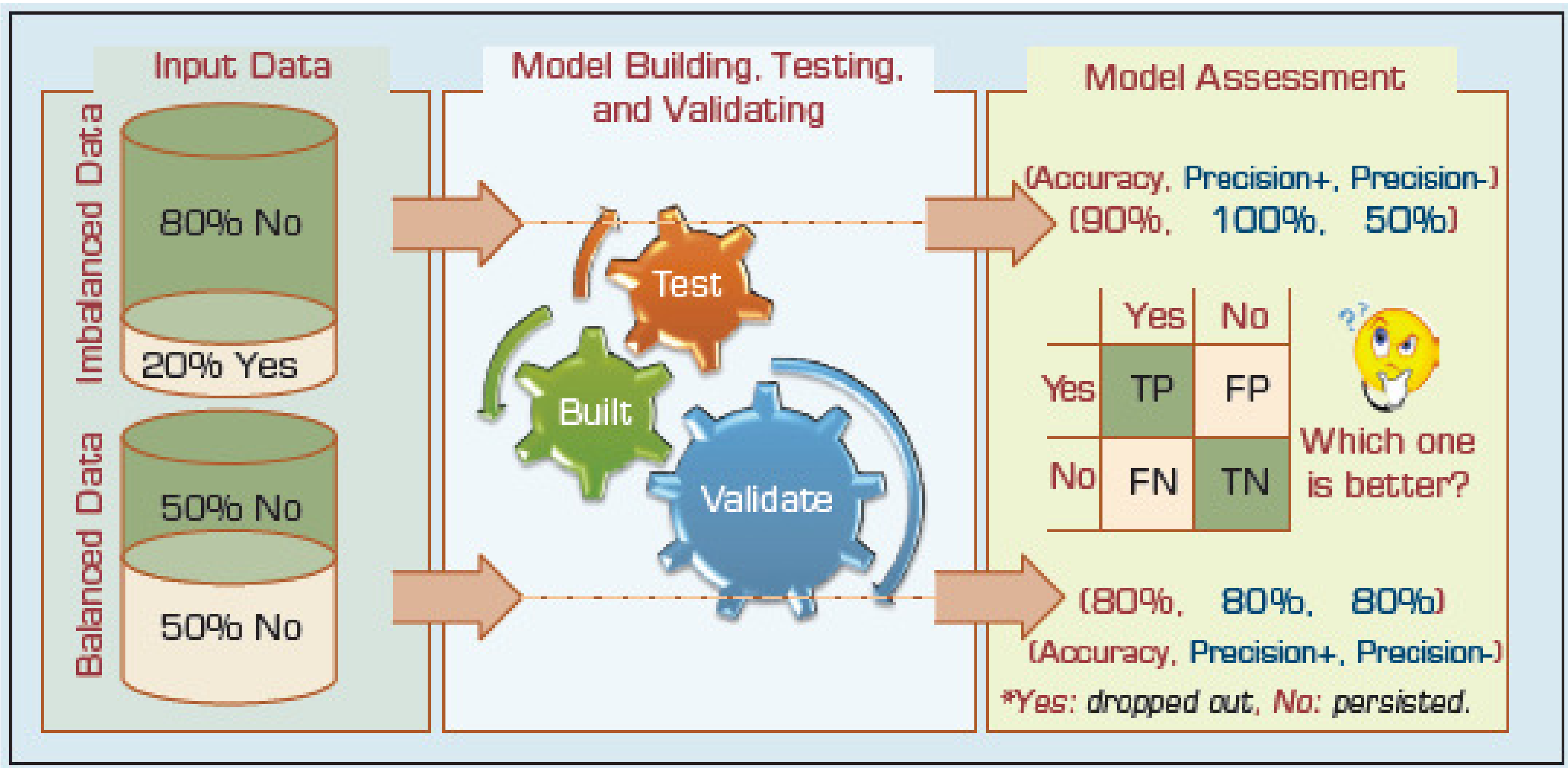
- Student retention
  - Freshmen class
- Why it is important?
- What are the common techniques to deal with student attrition?
- Analytics versus theoretical approaches to student retention problem



Application Case 2.2 (3 of 4)

Improving Student Retention with Data-Driven Analytics

- Data imbalance problem



Application Case 2.2 (4 of 4)
Improving Student Retention with Data-Driven Analytics

- Results…

TABLE 2.2 Prediction Results for the Original/Unbalanced Dataset

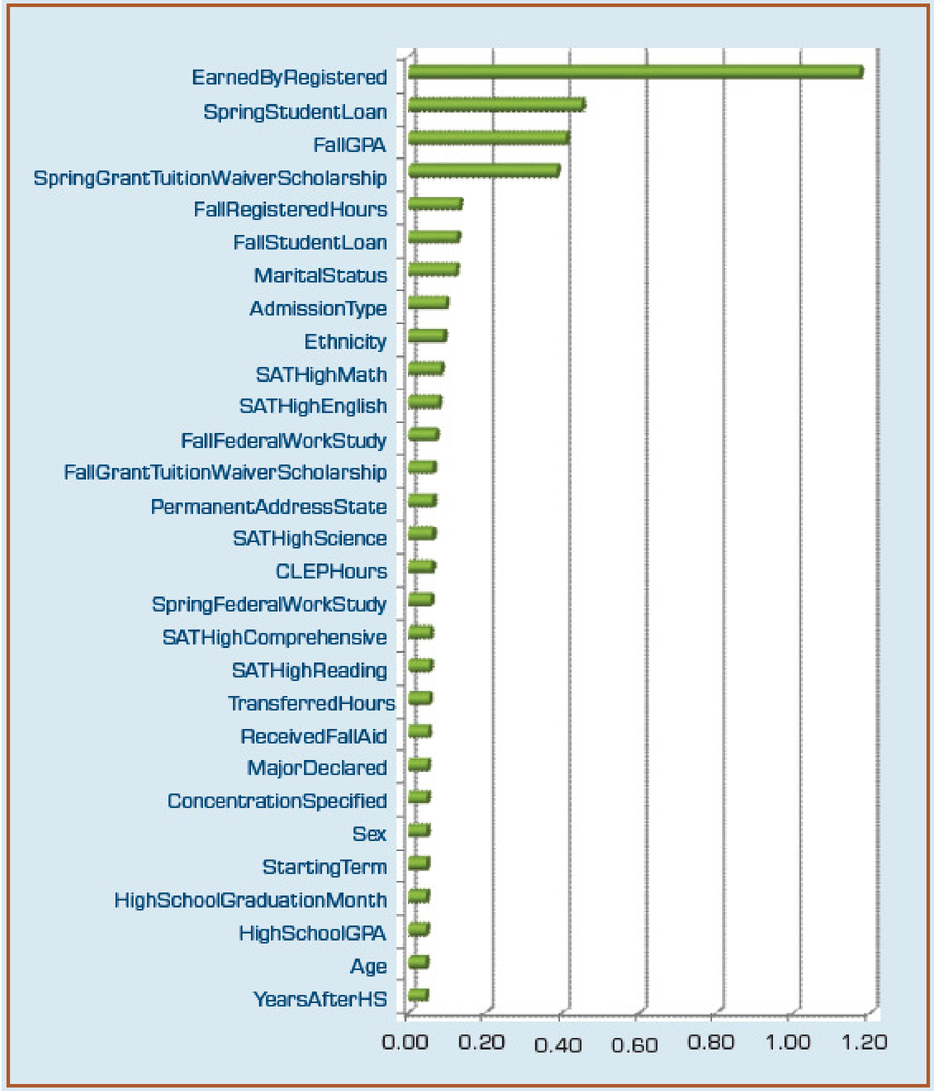
	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1494	384	1518	304	1478	255	1438	376
Yes	1596	11142	1572	11222	1612	11271	1652	11150
SUM	3090	11526	3090	11526	3090	11526	3090	11526
Per-Class Accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall Accuracy	86.45%		87.16%		87.23%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

TABLE 2.3 Prediction Results for the Balanced Data Set

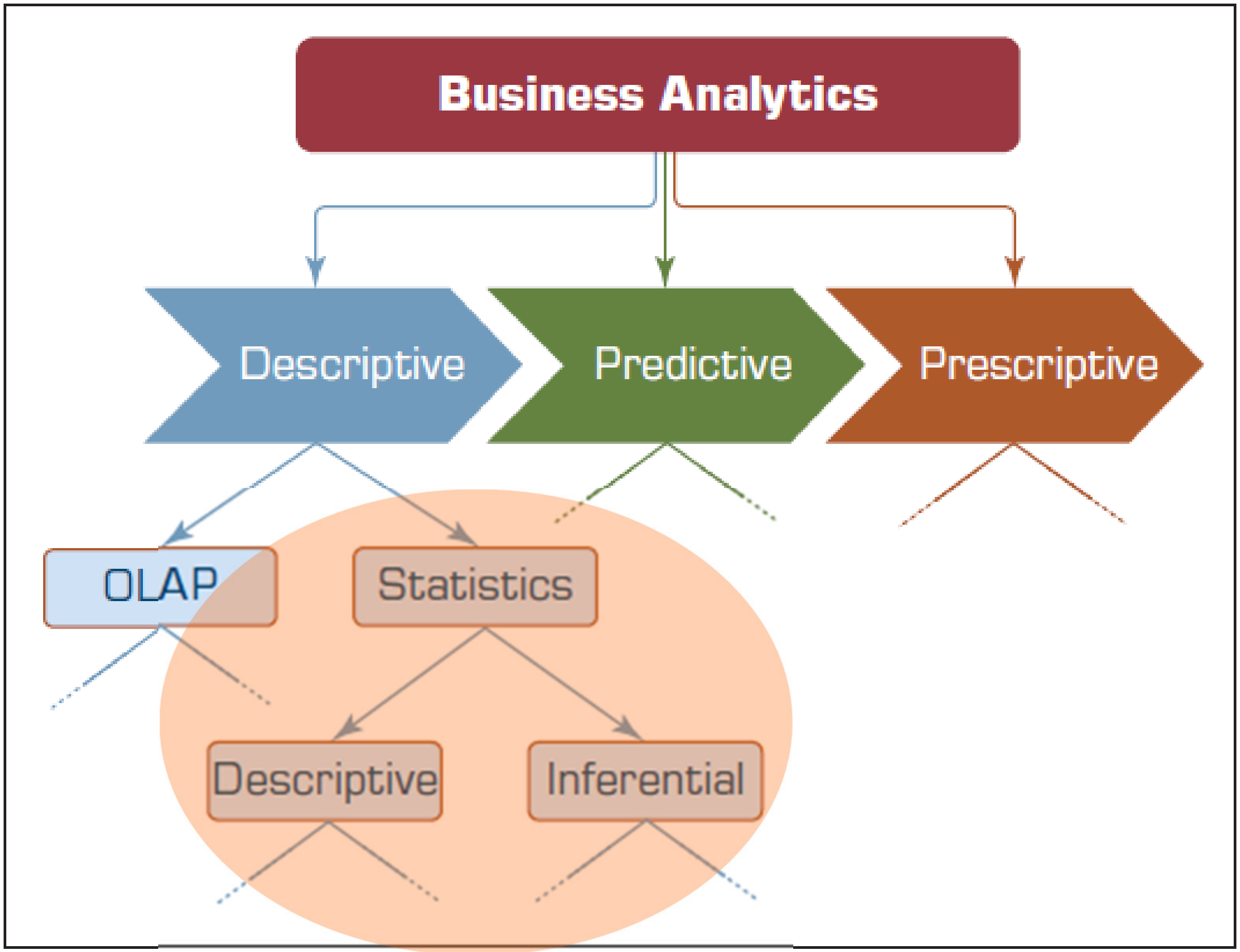
Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2309	464	2311	417	2313	386	2125	626
Yes	781	2626	779	2673	777	2704	965	2464
SUM	3090	3090	3090	3090	3090	3090	3090	3090
Per-class Accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall Accuracy	79.85%		80.65%		81.18%		74.26%	

TABLE 2.4 Prediction Results for the Three Ensemble Models

	Boosting (Boosted Trees)		Bagging (Random Forest)		Information Fusion (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%		82.10%	



Statistical Modeling for Business Analytics



Statistical Modeling for Business Analytics

- 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

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

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

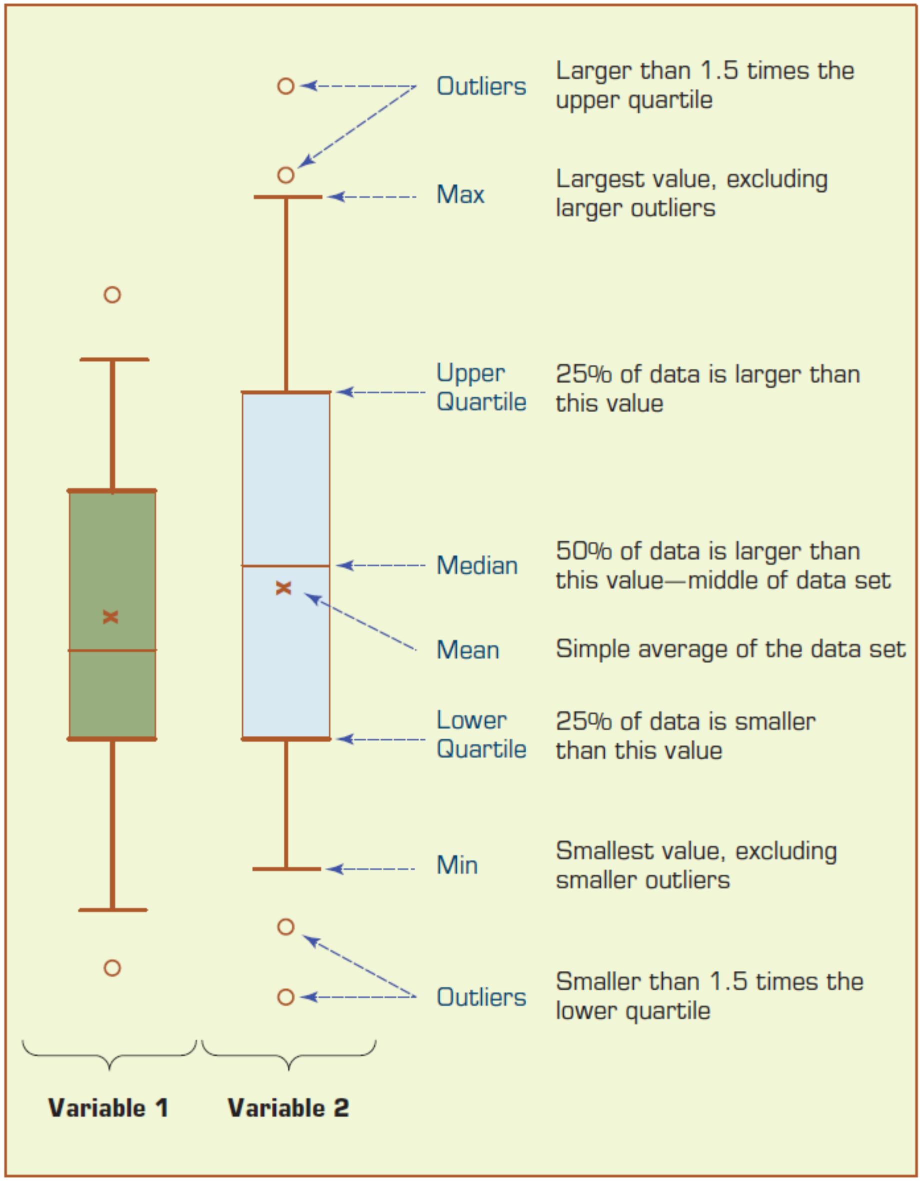
Standard Deviation

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

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

Descriptive Statistics Measures of Dispersion

- Quartiles
- Box-and-Whiskers Plot
  - a.k.a. box-plot
  - Versatile / informative



Descriptive Statistics
Shape of a Distribution

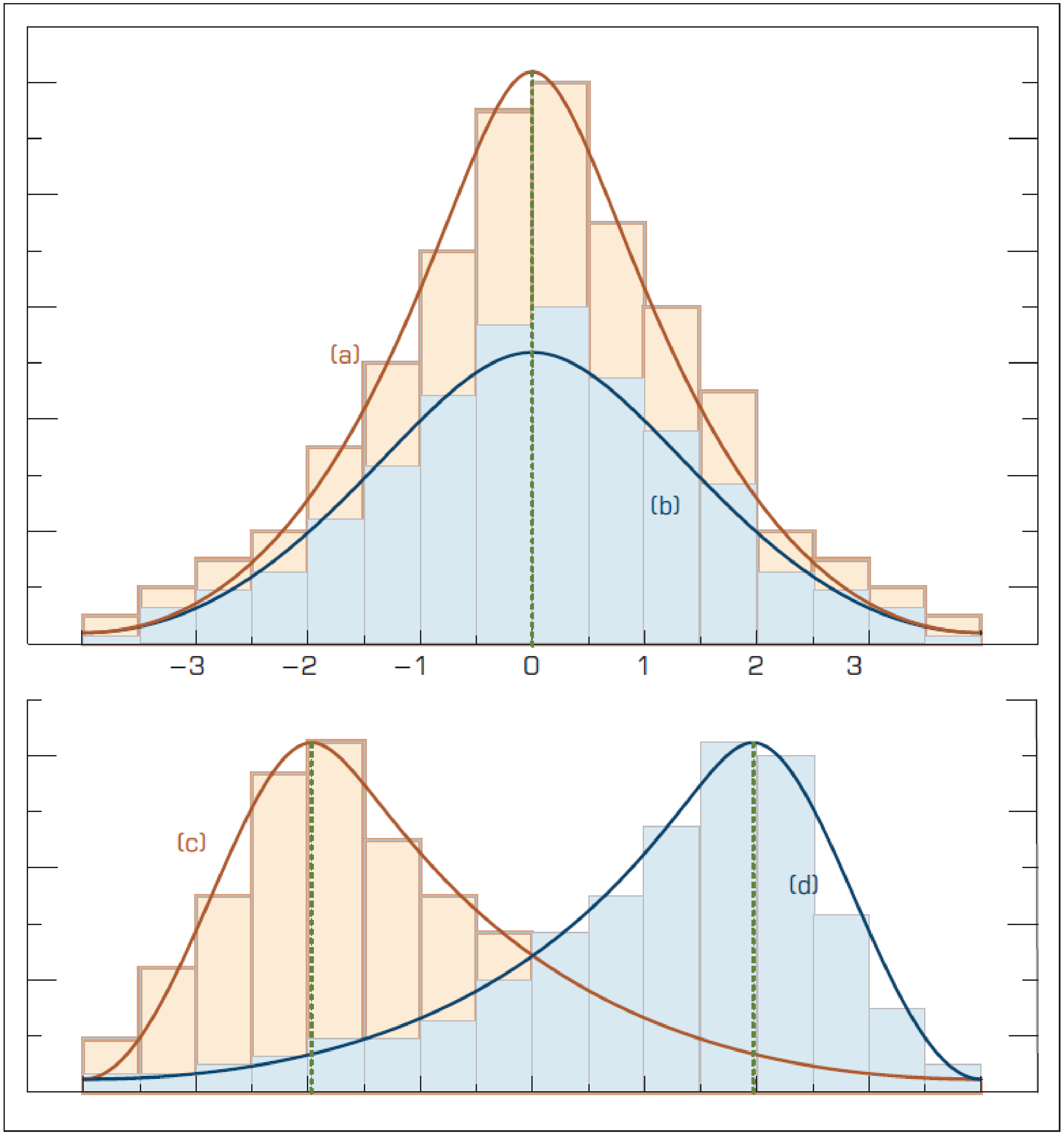
- Histogram – frequency chart
- Skewness
  - Measure of asymmetry

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

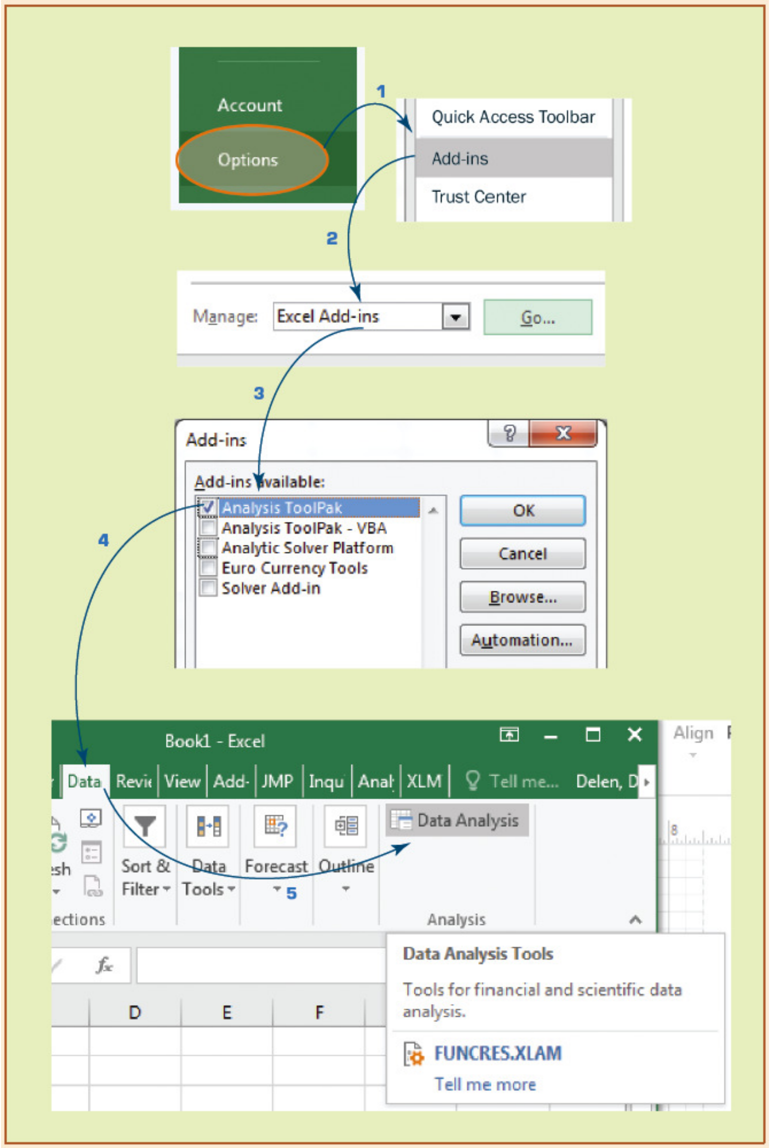
- Kurtosis
  - Peak/tall/skinny nature of the distribution

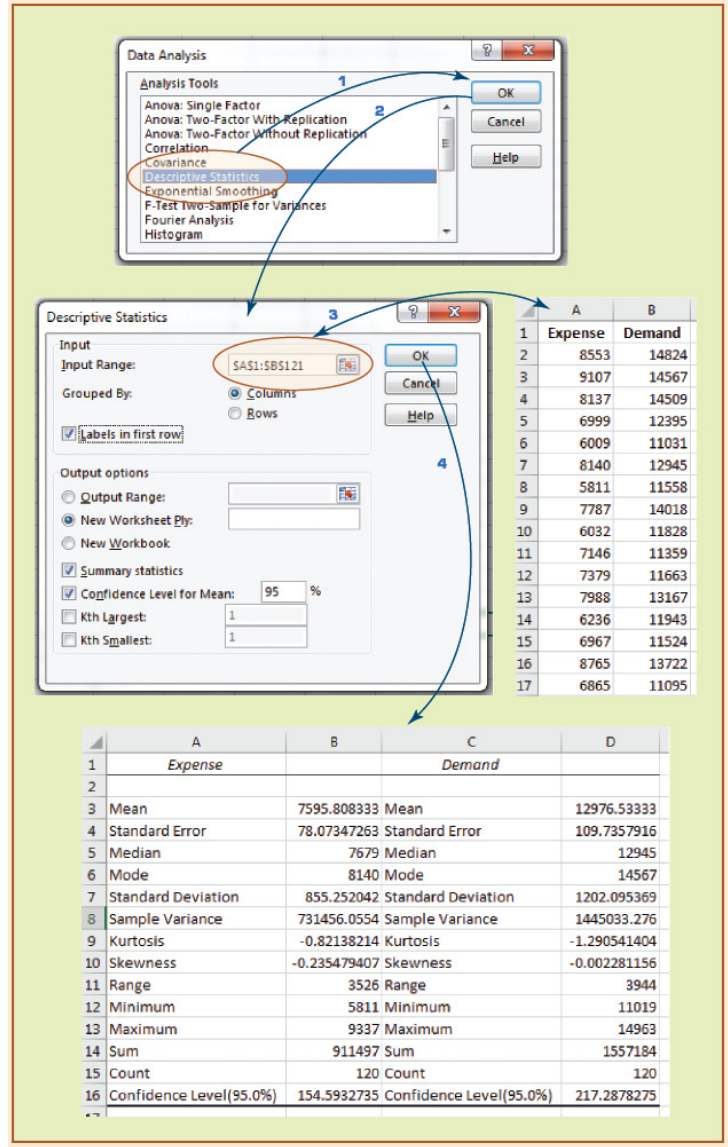
$$Kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

Relationship Between Dispersion and Shape Properties



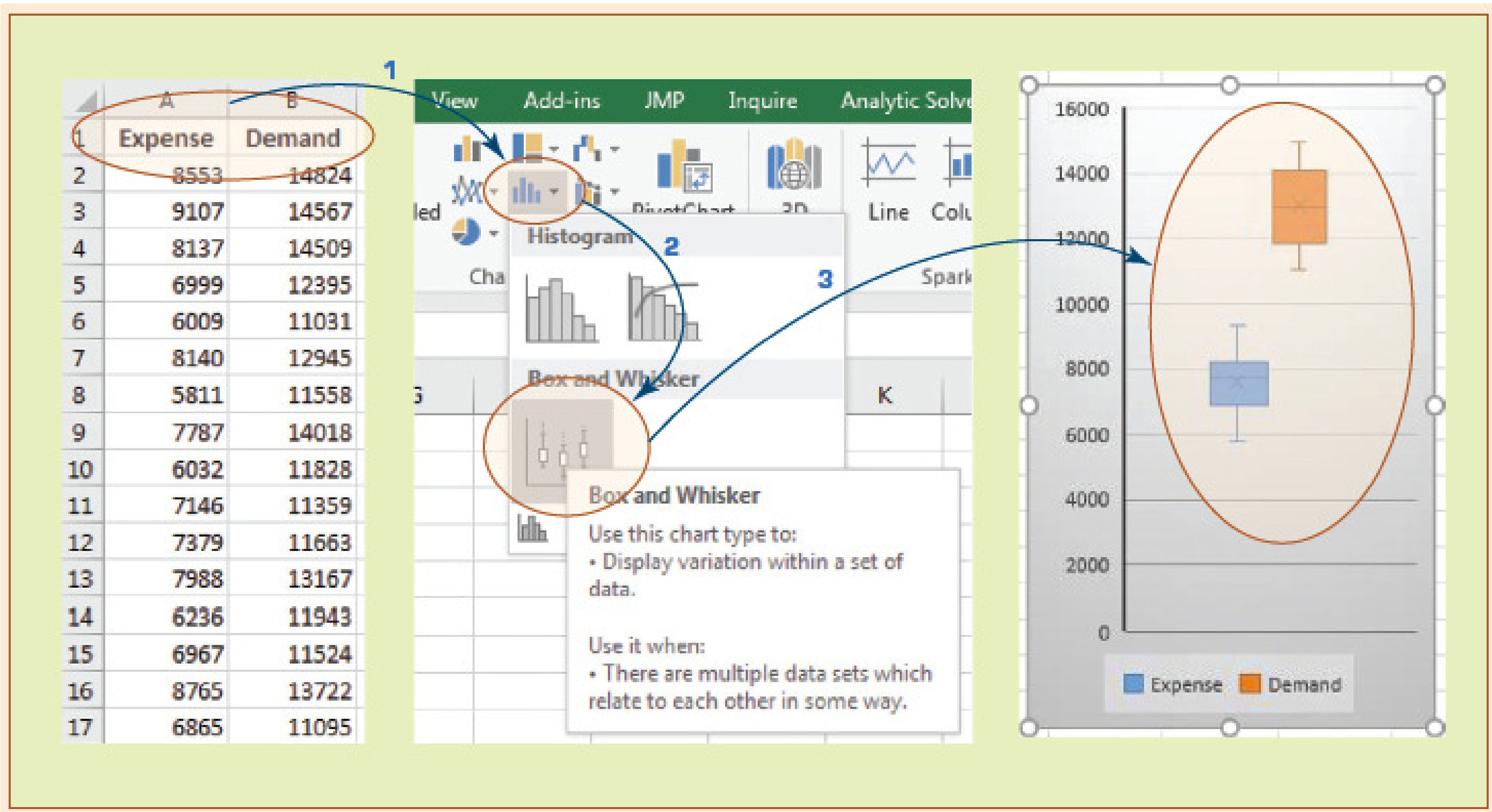
Technology Insights 2.1 – Descriptive Statistics in Excel





Technology Insights 2.1 – 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?

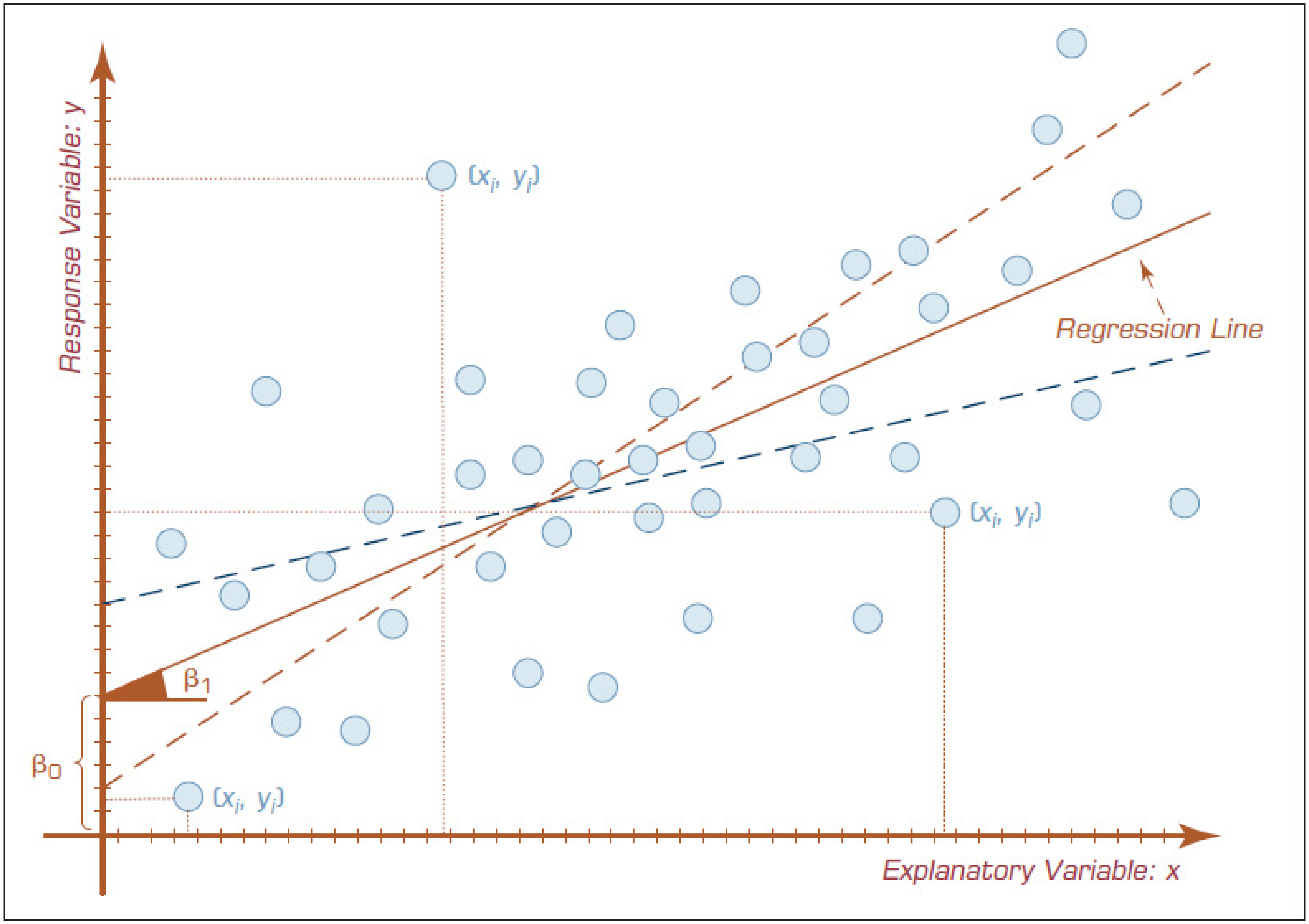
Regression Modeling for Inferential Statistics

- 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

- 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



Regression Modeling

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

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

- Multiple Linear Regression

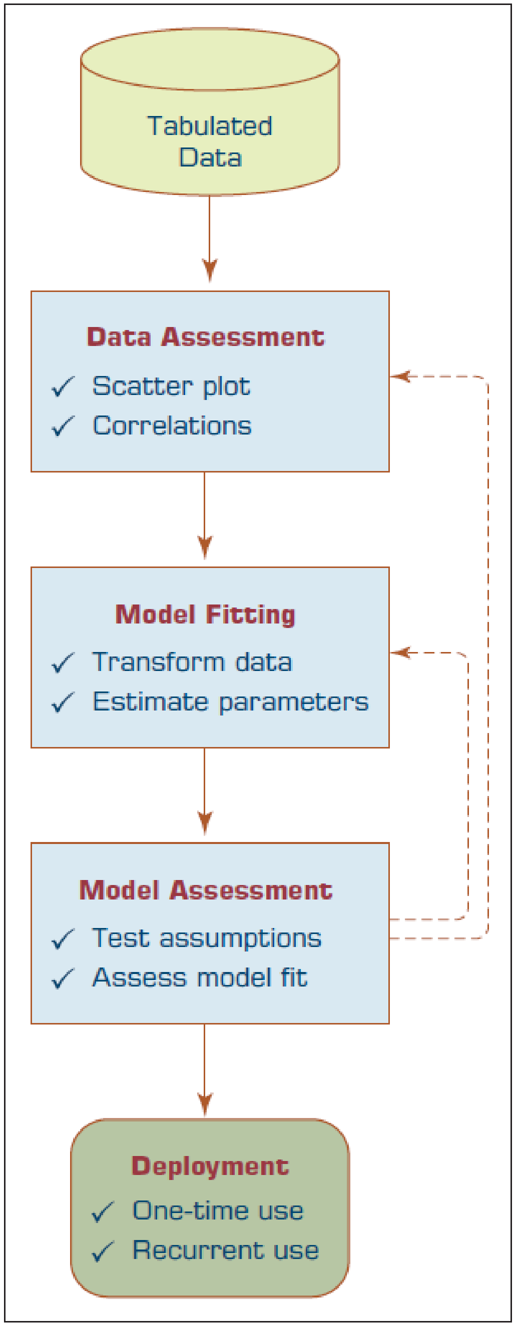
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

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

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



Regression Modeling Assumptions

- Linearity
- Independence
- Normality (Normal Distribution)
- Constant Variance
- Multicollinearity

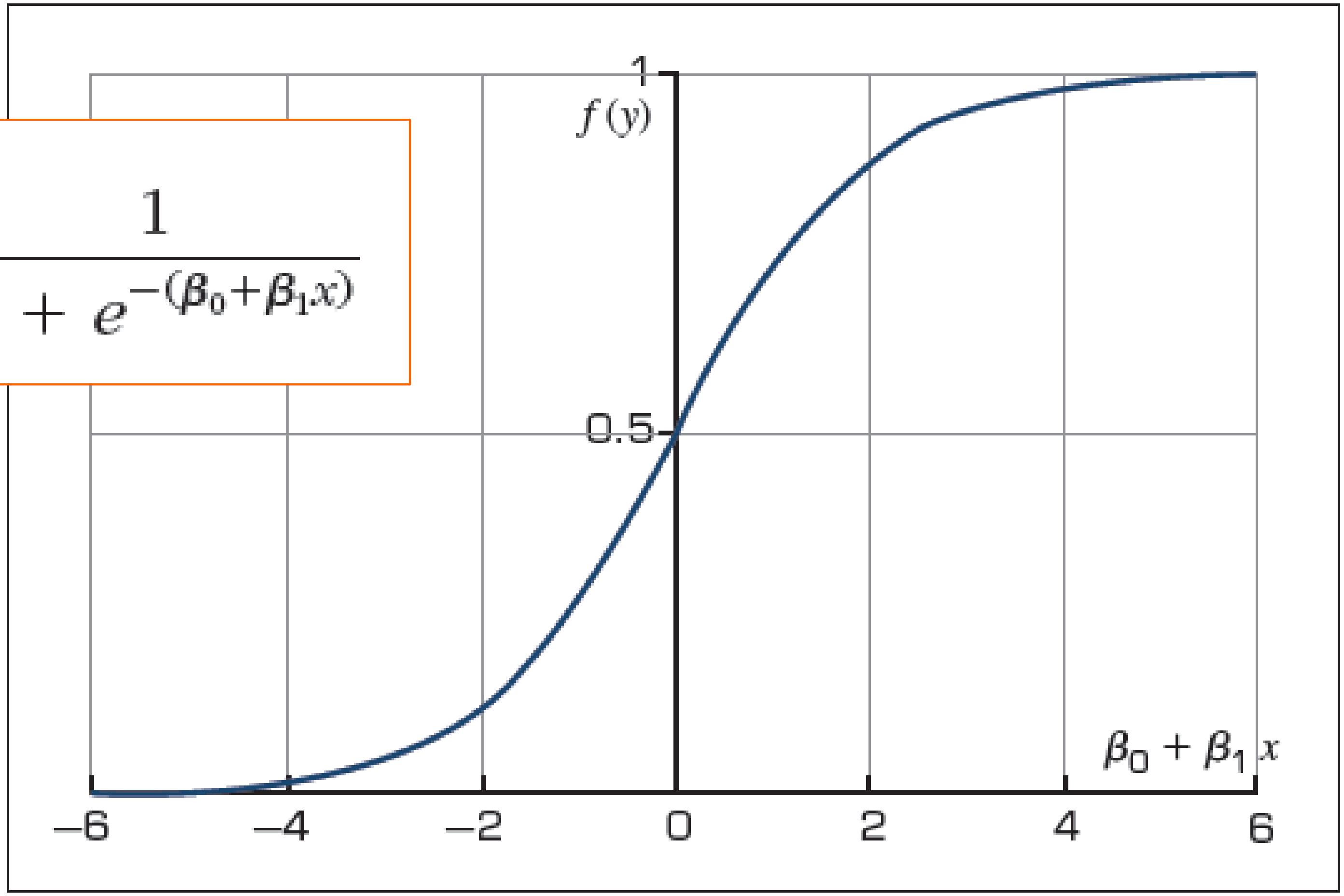
- What happens if the assumptions do NOT hold?
  - What do we do then?

Logistic Regression Modeling

- 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

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



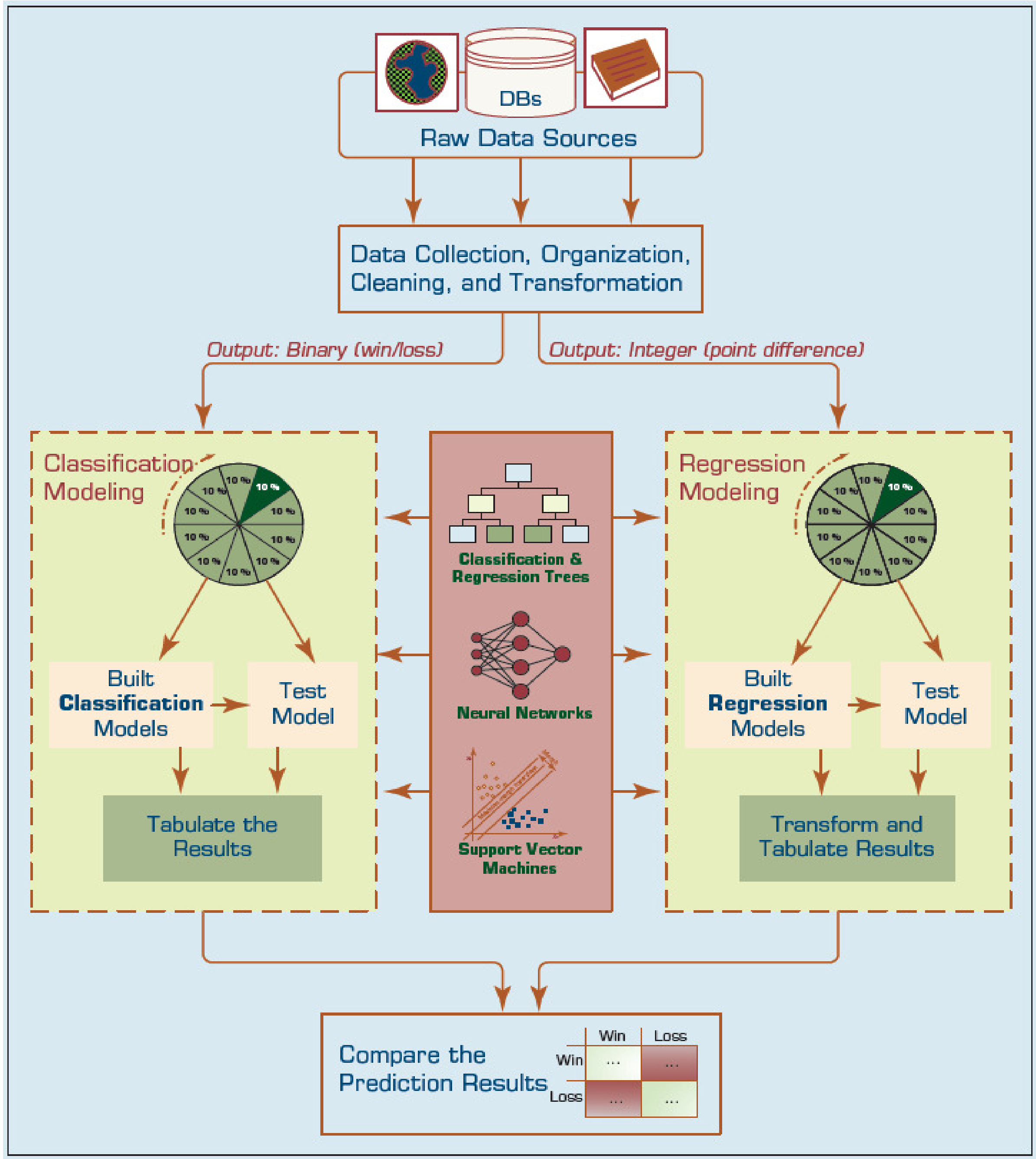
Application Case 2.4 (1 of 4)

Predicting NCAA Bowl Game Outcomes

Application Case 2.4 (2 of 4)

Predicting NCAA Bowl Game Outcomes

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



Application Case 2.4 (3 of 4)

Predicting NCAA Bowl Game Outcomes

Prediction Results

1. Classification
2. Regression

TABLE 2.6 Prediction Results for the Direct Classification Methodology

Prediction Method (Classification*)		Confusion Matrix Win	Confusion Matrix Loss	Accuracy** (in %)	Sensitivity (in %)	Specificity (in %)
ANN (MLP)	Win	92	42	75.00	68.66	82.73
	Loss	19	91			
SVM (RBF)	Win	105	29	79.51	78.36	80.91
	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 (** $p < 0.01$).*

TABLE 2.7 Prediction Results for the Regression-Based Classification Methodology

Prediction Method (Regression-Based*)		Confusion Matrix Win	Confusion Matrix Loss	Accuracy**	Sensitivity	Specificity
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 (4 of 4)
Predicting NCAA Bowl Game Outcomes

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?