Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.

This page intentionally left blank.

TDWI Predictive Analytics Fundamentals

TDWI takes pride in the educational soundness and technical accuracy of all of our courses. Please send us your comments—we'd like to hear from you. Address your feedback to:

info@tdwi.org

Publication Date: January 2018

TABLE OF CONTENTS

COURSE OBJECTIVES

You will learn:

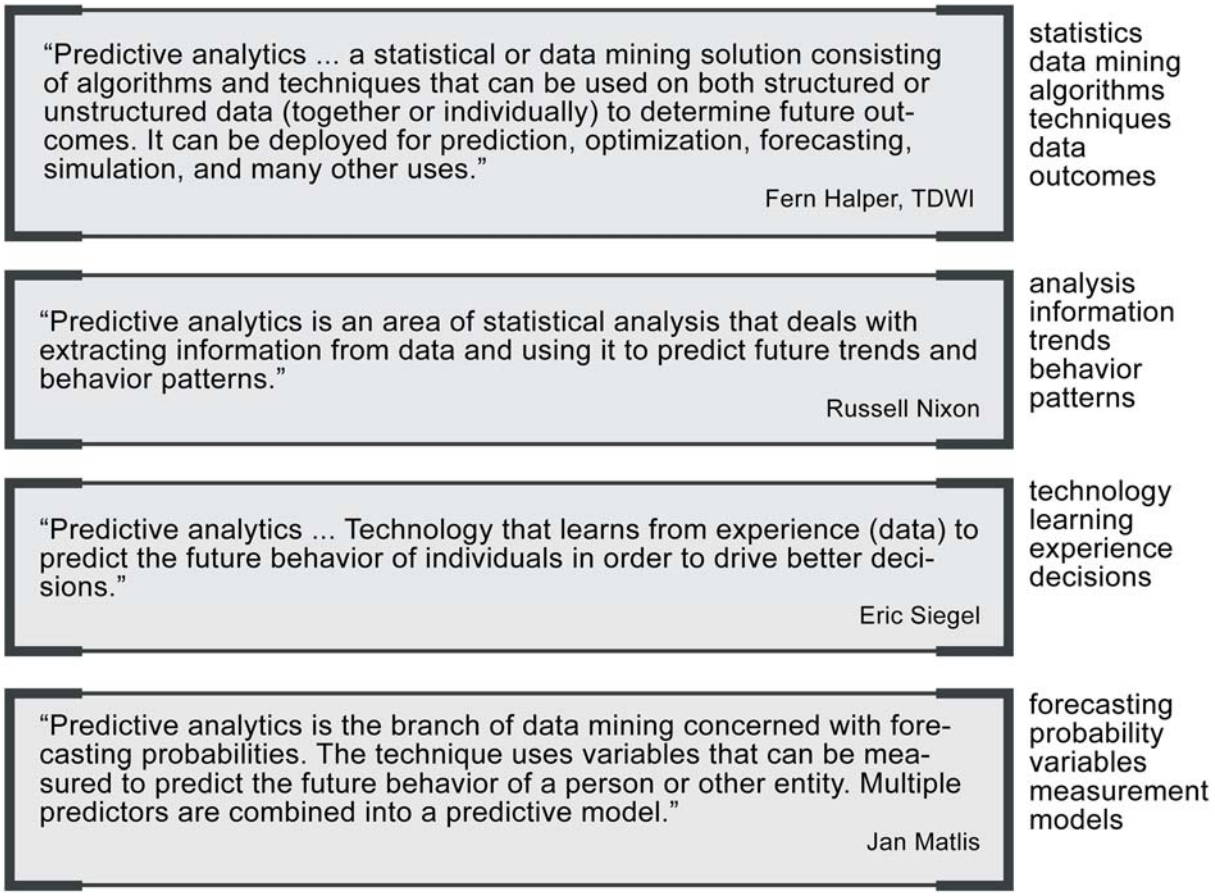
- ✓ ***Definitions, concepts, and terminology of predictive analytics***
- ✓ ***How predictive analytics relates to data science and BI programs***
- ✓ ***Structure, categories, and applications of predictive models***
- ✓ ***Enabling methods adapted from statistics, data mining, and machine learning***
- ✓ ***Proven development and implementation approaches***
- ✓ ***How human and organizational factors including team composition, structure, culture, collaboration, and accountabilities enable success***
- ✓ ***Why business, technical, and management skills are essential for success***
- ✓ ***Practical guidance for getting started with predictive analytics***

Module 1

Predictive Analytics Concepts

What and Why of Predictive Analytics

Predictive Analytics Defined



What and Why of Predictive Analytics

Predictive Analytics Defined

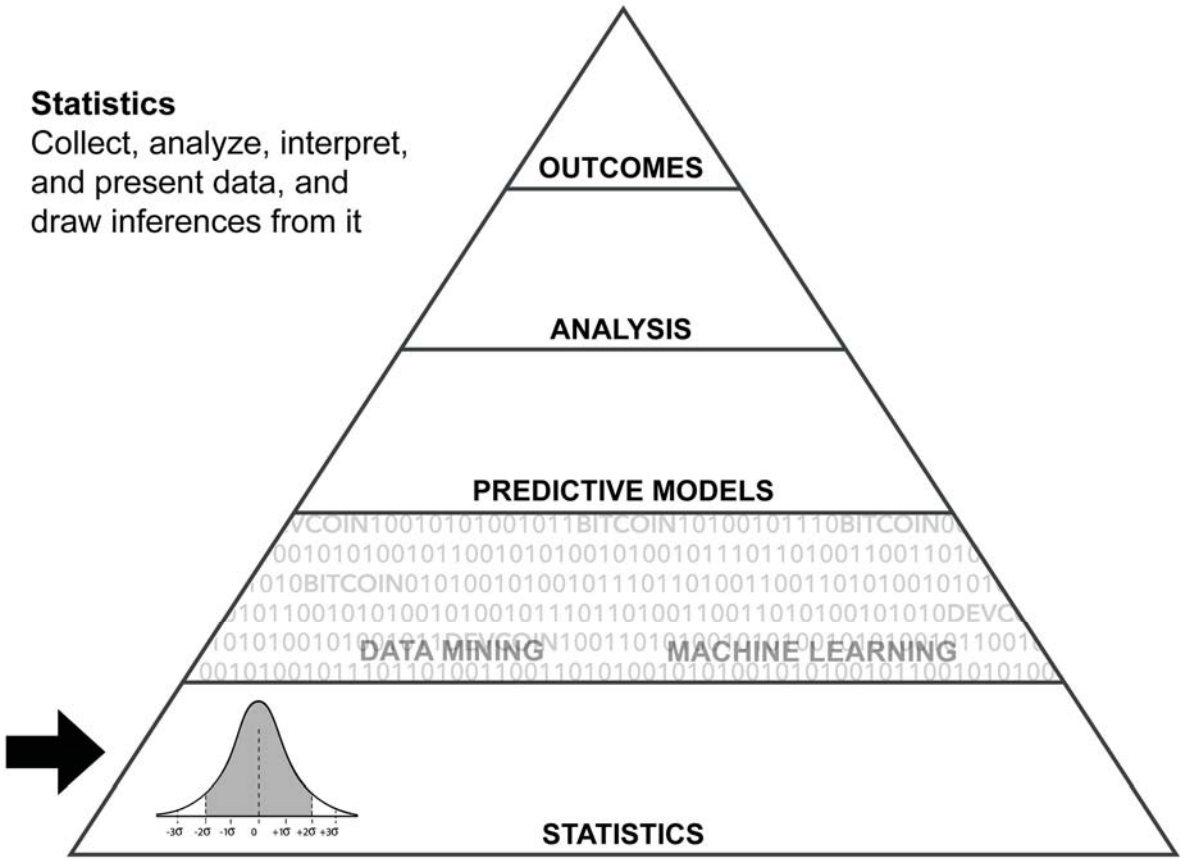
MANY DEFINITIONS The facing page illustrates four different but complementary definitions of predictive analytics. Each definition views predictive analytics from different perspectives—a solution, an area of analysis, technology that learns, and a branch of data mining. Combining the four definitions yields a well-rounded description of predictive analytics.

MANY ASPECTS Extracting key words from each definition provides a sense of the many different aspects of predictive analytics and the complexities inherent in the subject. Predictive analytics includes each of the following:

- Statistics—the foundation of data mining and data analysis
- Data mining—prerequisite technology, skills, and processes for predictive analytics
- Algorithms—specific mathematical and statistical methods applied in data mining
- Techniques—goal-oriented uses of data mining algorithms
- Data—the raw material of analytics
- Outcomes—business results, impact of decisions and actions
- Analysis—finding meaning behind data and statistics
- Information—data with context, meaning, and purpose
- Trends—general directional movement of outcomes and general behavioral tendencies
- Behavior—action, reaction, or response to specific circumstances
- Patterns—typical or normative ways of acting and responding
- Technology—consisting mostly of software designed for data mining and predictive analytics
- Learning—specifically, machine learning where computers acquire knowledge by examining data
- Experience—past events and outcomes that are encoded as data
- Decisions—the choices that influence future outcomes
- Forecasting—estimating future outcomes by understanding the indicators inherent in the past (experience)
- Probability—a measure of how likely it is that something will occur
- Variables—specific data items that are the subjects of analysis
- Measurement—quantification, the assignment of numbers to represent properties of something
- Models—application of algorithms to generate predictions and make inferences about relationships

The Foundation for Predictive Analytics

Statistical Foundation



The Foundation for Predictive Analytics

Statistical Foundation

THE BASICS OF DATA ANALYSIS

The basis of predictive analytics begins with statistics—the most basic set of techniques for data analysis that make it possible to segment populations, identify behavioral trends and patterns, and forecast and quantify future behaviors.

The topic of statistics is covered in more depth in Module 2 of the course.

Predictive Analytics in BI Programs

Predictive Analytics in the BI Stack

Decision Management	business rules engines, optimization, simulation and forecasting
Business Analytics	mining, modeling, visualization, **predictive analytics**, text analytics, geospatial analytics
Business Applications	enterprise reporting, performance mgmt, scorecards, dashboards, operational BI
Information Services	query and access, reporting (tabular and graphical), OLAP
Data Integration	ETL/ELT, data virtualization, big data integration
Data Management	data storage, DBMS, big data technologies, data profiling, data quality, metadata management
Data Sourcing	internal source systems; data connectivity and APIs; syndicated and subscription data services
Infrastructure	servers, operating systems, networks, security, performance

Predictive Analytics in BI Programs

Predictive Analytics in the BI Stack

PREDICTIVE ANALYTICS AND BUSINESS INTELLIGENCE

The business intelligence (BI) stack is a layered view of the components of BI from infrastructure to decision management. Dependency can be viewed up and down the stack. Each layer depends upon the layers below and provides support for the layers above.

Predictive analytics specifically and business analytics in general are often positioned as parts of a BI program. The facing page illustrates placement of predictive analytics in a BI stack, as part of the business analytics layer. It depends upon all of the infrastructure and data management layers and may be used to support decision management applications.

Becoming Analytics Driven

Business Driven

Becoming Analytics Driven

Business Driven

BASED ON STRATEGIC INTENT

Organizations that recognize the opportunities for becoming analytics driven as part of their competitive positioning embrace the need for predictive analytics at the senior leadership level.

It is common that the organization's strategy is built around capabilities that are enabled by a wide range of analytics techniques. In this scenario, predictive analytics capabilities are embedded in the firm's strategy as a key enabler.

The approach to this type of organizational transformation is built around analytics and is led by a senior leadership team. It is common to see a new C-level executive called the chief analytics officer (CAO), the chief data officer (CDO), or a chief strategy officer (CSO) to champion and lead this transformation.

An LOB (line of business) analytics leader provides direction and coordination at a business segment level. He or she operates at a level lower than the enterprise but is accountable for implementing analytics capabilities within a defined area using a strategic perspective.

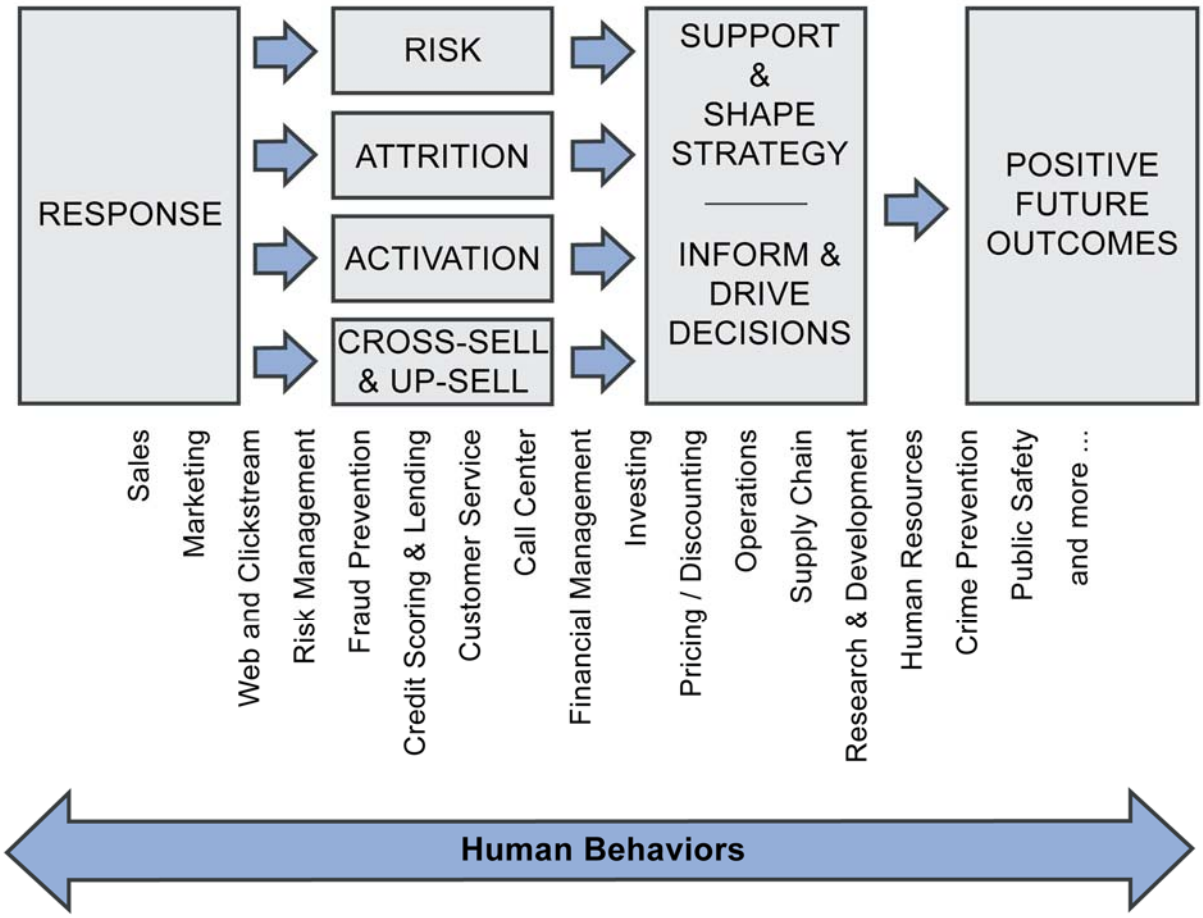
To be effective, the organization must understand and change how decisions are made and how employees are incentivized to become data driven in their job functions.

Functional areas within the firm that are expected to drive new business opportunities are identified and targeted as high-leverage places for implementing new predictive capabilities.

Areas such as supply chain management, marketing, customer care, quality management, risk management, and talent management are examples of potential areas could be targeted for the transformation.

Common Applications for Predictive Analytics

What Businesses Need to Predict



Common Applications for Predictive Analytics

What Businesses Need to Predict

RESPONSE PREDICTION

Predictive analytics is predicated on the concept of predicting human behaviors—what people will do in specific circumstances. Every predictive analytics project begins with response prediction where the goal is to understand how people (and various segments of a population) will respond to a specific situation or stimulus.

EXTENDING THE RESPONSE MODEL

Response predictions are typically extended or adapted to specific needs and circumstances. A response model may be extended to predict:

- Risk—predictions about segments of a population that may engage in fraud, commit crimes, compromise workplace safety, etc.
- Attrition—predictions about segments of a population that may be lost as customers, employees, contributors, partners, etc.
- Activation—predictions of probability (by segment) to set a process in motion, such as activating a trial version of a software product
- Cross-sell and up-sell—predictions of probability that purchasers will respond to suggestions for related products and services

APPLYING THE RESPONSE MODEL

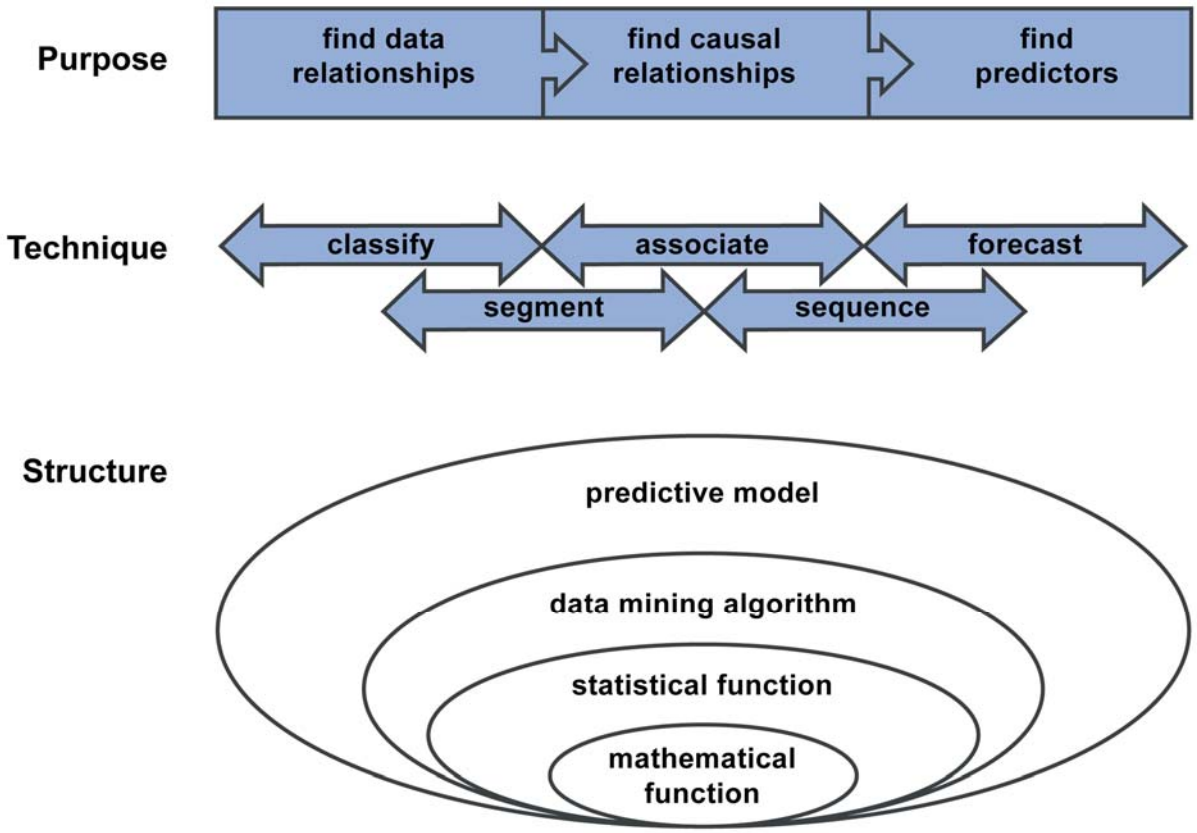
As already discussed, the value of predictive analytics is achieved by applying the predictions to support and shape strategy and to inform and drive decisions.

BUSINESS OUTCOMES

Predictions of response, extended to context of business need and used to drive positive business outcomes, are the keys to effective predictive analytics. Positive business outcomes are specifically related to business domains. The facing page illustrates many of the common business domains—from sales and marketing to public safety—where value is created with predictive analytics.

The Language of Predictive Analytics

Making Sense of the Terminology



The Language of Predictive Analytics

Making Sense of the Terminology

COMMON WORDS WITH SPECIFIC MEANINGS

The language of data mining and predictive analytics can be confusing to those who are just becoming familiar with the field. In many cases commonly used words have very specific meanings in data mining and predictive modeling contexts.

PURPOSE

A data mining model typically serves to find one of three things:

- Data relationships—hidden patterns in the data that are useful for gaining new knowledge and understanding
- Causal relationships—relationships that indicate cause and effect
- Predictors—the variables in a relationship that are useful to predict future outcomes.

TECHNIQUE AND STRUCTURE

Data mining uses multiple techniques to achieve different objectives. These techniques, discussed in greater depth later in this course:

- Classify—find groups of similar objects in a set of data
- Segment—divide data into subsets based on the groups found through classification
- Associate—find relationships among variables (attributes) in a set of data
- Sequence—discover patterns in the data where sequence or order of occurrence is apparent across multiple events
- Forecast—use historical data to understand future trends and probabilities

STRUCTURE

A data mining model is a combination of data and software that is used to draw inferences from data relationships and to generate predictions about the subjects of the data. The software applies algorithms to data to implement selected data mining techniques.

- A *model* is built using one or more algorithms.
- An *algorithm* applies one or more functions to determine a result.
- A *statistical function* applies one or more mathematical functions to derive statistical values. Some statistical functions called *distributions* generate a range of possible values for a given input value. These are discussed in more detail later in the course.
- A *mathematical function* is an equation or formula that generates an output value based on or more input values.

Module 2

Models and Statistics

Predictive Models

What Are Models?

Models

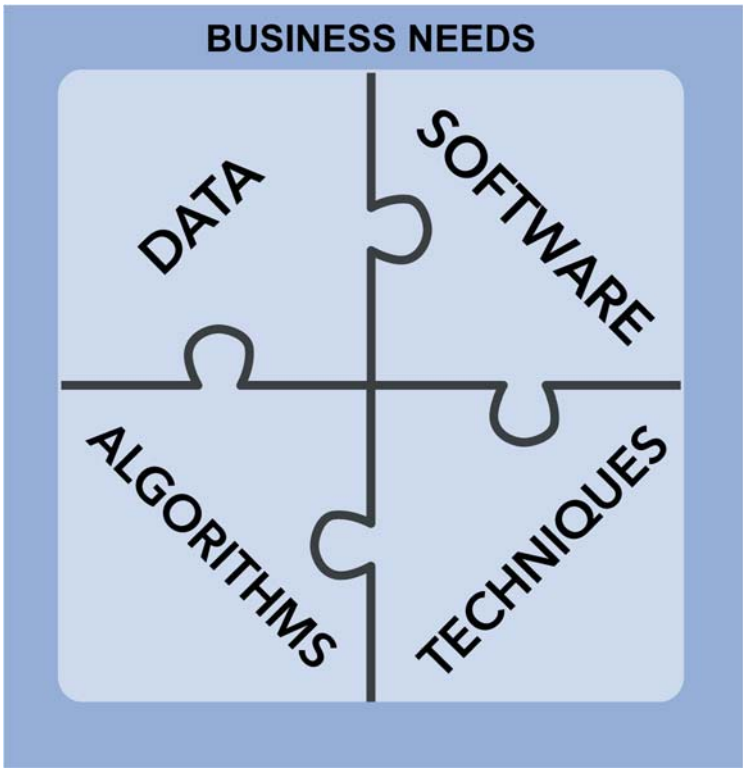
- Simplified representations of reality that meet defined information needs

Main Components of Models

- Data
- Software
- Algorithms
- Techniques

Other Components of Models

- Symbols
- Rules
- Diagrams
- Physical items



Predictive Models

What Are Models?

MODELS

Models are simplified abstractions of reality. By design, models are simplified to a certain level of detail that still maintains their usefulness and relevance to a given problem while eliminating detail that is not relevant.

An *analytical model* combines data with software to help draw inferences from the data and to create predictions for some of the attributes contained in the data. The software applies algorithms to the data to implement selected data mining, statistical, or machine learning techniques.

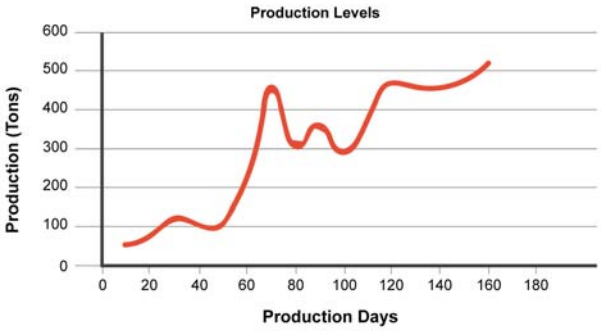
The selection of data, algorithm, and technique for a given situation is driven by the problem context and by the needs people have for the model's informational output. Models are the building blocks used by automated processes that search for patterns and relationships in large data sets.

Descriptive Statistics

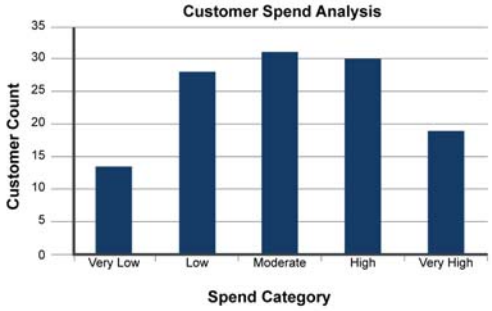
Variables

	DISCRETE	CONTINUOUS
DEFINITION	A data item with a finite number of possible values	A data item that can have any value within a range of real numbers
EXAMPLES	• Children in a family • Number of defects • Survey scale response	• Household income • Weight of an individual • Rainfall in a location
VALUES	Integer	Real numbers
SCALES	Ordinal Nominal	Ratio Interval
DATA TYPES	Quantitative Counting	Quantitative Measurement
PRIMARY USE	Categorize	Calculate

Continuous Variables



Discrete Variables



Descriptive Statistics

Variables

DESCRIBING WHAT YOU KNOW

The discipline of predictive analytics is rooted in statistics, a mathematical discipline that describes the world through data. The most basic statistical concepts are *descriptive*; they describe what is known about a data set, without making any generalizations to the larger world.

DATA AS VARIABLES

Variables are the fundamental building blocks of statistics. A *variable* is any data element that describes something in the world. It is called a variable because the value may vary between members in a population. Variables may be either qualitative or quantitative. Although statistics is a mathematical discipline, not all variables are numeric, nor are all numeric variables quantitative.

DISCRETE AND CONTINUOUS VARIABLES

Quantitative variables are further classified as discrete or continuous.

A *discrete value* is one that cannot take on all values within the limits of its range of possible values. A single family, for example, may have 2 children or 3 children but they can't have 2.7 children. Discrete variables are constrained to be integers.

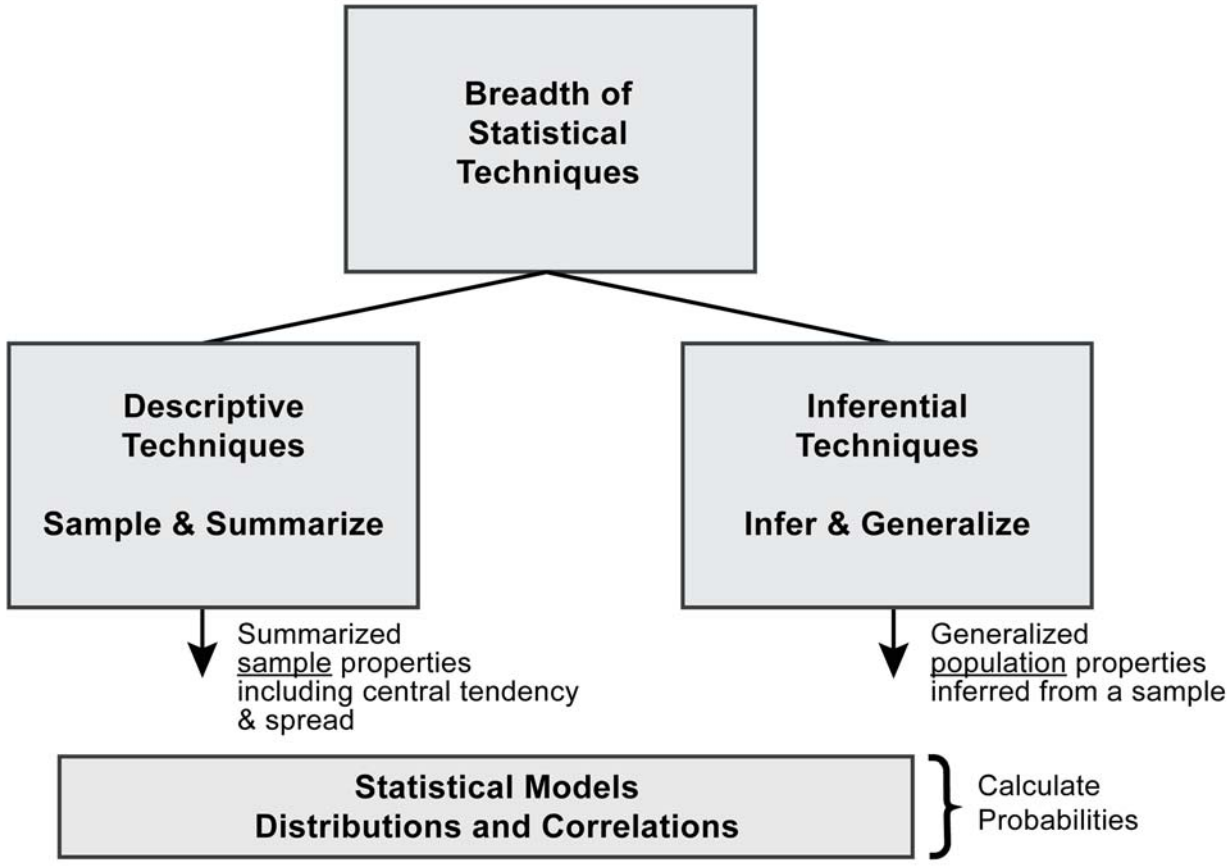
A *continuous variable* can take on any value within the limits of its range of possible values. A person's weight may be 174 pounds or 175 pounds, but it may also be 174.66 pounds. Continuous variables are real numbers and can have fractional values.

The table at the top of the facing page summarizes the differences between discrete and continuous variables.

The two graphs on the facing page show examples of continuous and discrete variables showing real and integer values respectively.

Inferential Statistics

Modeling the Population



Inferential Statistics

Modeling the Population

SAMPLE AND SUMMARIZE

As you have seen, descriptive statistics is used to summarize a data set, describing the shape of the data.

The shape of the sample data can be described by the following descriptive concepts.

- Central tendency statistics (mean, median, and mode)
- Spread statistics (variance, range, and standard deviation)
- Skew statistics (skewness)
- Correlation (regression)

These descriptive characteristics are directly calculated from the data sample. They describe the shape of the sample data.

INFER AND GENERALIZE

Inferential statistics use the sample data to describe the larger population. Inferential techniques help us draw conclusions from a sample that can be generalized to the overall population.

STATISTICAL MODELS

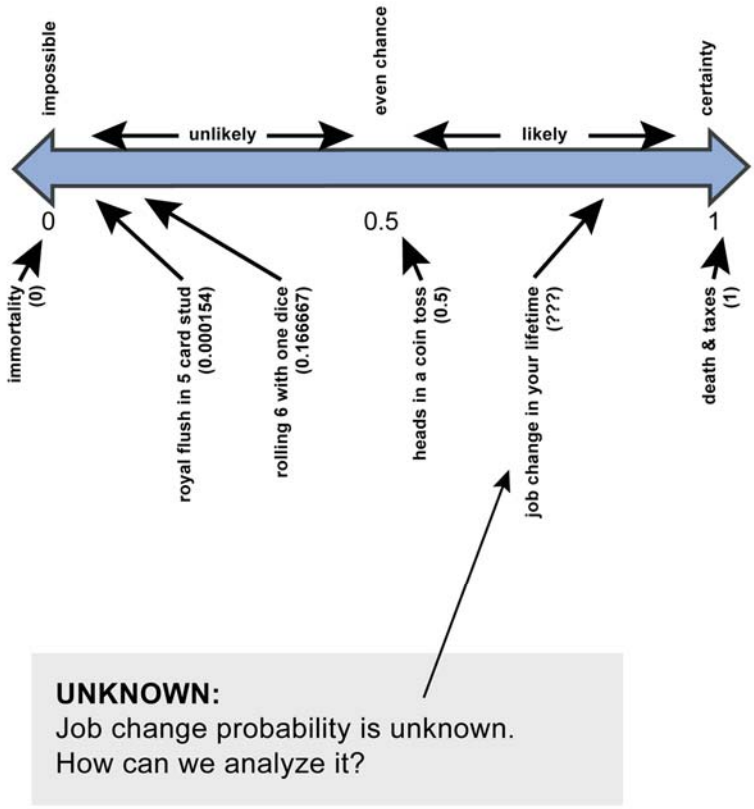
Statistical techniques from both the descriptive and inferential branches are used to create models. These models use data to describe distribution patterns, shapes, and correlations. Unlike descriptive models, these models are intended to describe the world beyond the sample data.

Examples of statistical models that describe the larger population include probability distributions, which describe the expected spread of values, and regression models, which describe expected correlations. Probability distributions will be discussed later in this module. Regression is described in *Module 3: Regression Model Examples.*

Statistical models are the simplest form of predictive models. Additional techniques and algorithms are explored in *Module 4: Building Predictive Models.*

Probability

Estimating Likelihood



UNKNOWN:
Job change probability is unknown.
How can we analyze it?

Business Context
What are the goals of analysis?

Data
What data is needed?
What are the characteristics of the data?
What does the data tell us?
(e.g., is job change an independent or dependent event?)

Analytic Modeling
Which techniques and how to apply them?

Evaluation
Are the analysis goals met?
Have we measured probability of job change?
Is it a useful and reliable predictor?

Probability

Estimating Likelihood

MEASURING PROBABILITY

Probability is an important concept in predictive analytics. FICO, the decision management company best known for credit scoring, says that "predictive analytics turns uncertainty into usable probability." In statistical terms, probability measures how likely it is that something will occur. The value of the measure is always in the range of zero to one, where zero corresponds with impossible and one corresponds with complete certainty. The scale on the facing page illustrates several examples of probability between the two extremes.

ANALYZING PROBABILITY

Probability analysis is investigation and study to turn uncertainty into a usable probability measure. Many of the examples shown here are mathematically certain—probability of heads on a coin toss is always 0.5. One of the examples—job change—is uncertain and a good candidate for analysis. Statistical analysis and data mining provide the means to perform that analysis, beginning with problem context and ending with a useful probability measure.

PROBABILITY MODELS

There are many techniques that help us create models that calculate probabilities. Some of the terms you may encounter include:

- Statistical models
- Distribution models
- Data mining techniques
- Machine learning algorithms

In statistics, there are multiple ways to estimate probabilities. This module explores probability distribution models. Module 3 will explore regression models. In predictive analytics, additional techniques include data mining and machine learning. These will be discussed in Module 4.

Module 3

Regression Model Examples

Regression Models

Overview

Regression Models
Statistical models that estimate relationships between a dependent variable and one or more independent variables.

- Wikipedia

Relevance to Predictive Analytics

- Linear regression models can forecast future values of continuous dependent variables.
- Logistic regression models can estimate probabilities of future discrete dependent variables.
- They are common techniques for implementing predictive models.

Regression Models

Overview

REGRESSION MODELS

Regression models employ statistical techniques to estimate quantified relationships between a dependent variable and one or more independent variables. There are a wide variety of regression techniques, which describe relationships by fitting lines, curves, or shapes to the observed data. A simple form of regression model describes a relationship by specifying the formula for a straight line that best fits the data.

There are several types of regression models that are influenced by the types of variables in the data set.

LINEAR AND LOGISTIC

The following two types of regression models are commonly used in predictive analytics applications:

- *Linear* regression is used forecast values of continuous variables.
- *Logistic* regression is used to estimate probabilities of discrete events to classify those events according to a defined criterion for likelihood.

Linear Regression Models

Overview

Linear Regression Models
Statistical models that estimate relationships between a continuous dependent variable and one or more independent variables.

- Wikipedia

Areas of Applicability
- Use historical data to create models that forecast and predict values of continuous variables based on input independent variables.
- Single-input models are called simple linear regression models.
- Multi-input models are called multiple linear regression models.

Linear Regression Models

Overview

LINEAR REGRESSION

Linear regression models calculate continuous dependent variables to predict or forecast their future values.

A linear regression model is used when the dependent variable is continuous and the independent variables are continuous. The technique can also support ordinal and nominal (categorical) independent variables if they can be transformed to numerical values.

Linear regression models are created using statistical techniques. These techniques vary in how they arrive at a formula (or model) that best describes the observed data. Regression techniques attempt to minimize the error between the observed data and the generated model. The statistics and machine learning communities provide collaboration and contribution to development of algorithms in this area.

APPLICATIONS

Regression techniques use historical data to estimate the parameters of a model that may be used to forecast and predict future values of the dependent variable.

Simple linear regression models are developed to study the influence of a single independent variable on a single dependent variable. An example is provided on the following pages.

Multiple linear regression models may be developed to study the collective influence of several independent variables on a continuous dependent variable. When multiple independent variables are considered simultaneously, the interaction effects of the independent variables influencing the dependent variable can be studied and analyzed. This type of model helps managers make informed trade-off decisions by adjusting the decision variables in a manner that produces an acceptable output value.

Logistic Regression Models

Overview

Logistic Regression Models

Statistical models that estimate relationships between a binary dependent variable and one or more independent variables. Used for predicting and classifying outcomes.

- Wikipedia

Areas of Application

- Use historical data to create models that calculate probability values for a binary condition or event to occur
- Independent variables can be continuous, nominal (categorical) or ordinal
- Single-input models are called simple logistic regression models
- Multi-input models are called multiple logistic regression models

Logistic Regression Models

Overview

LOGISTIC REGRESSION

Logistic regression is used when the dependent variable is a binary category. Examples of binary categories include true vs. false, yes vs. no, win vs. lose, and so forth. Logistic regression produces a probability for the desired outcome based on the values of independent variables.

The independent variables can be continuous, ordinal, or nominal. When ordinal or nominal independent variables are used, they must be transformed to numerical values.

Logistic regression models use the logit function in producing output variables. Typically, the model produces the log of the odds ratio, then transforms it into a useful probability value.

The following pages provide a simple example of logistic regression with one independent variable. As with linear regression, there are multiple techniques for logistic regression. The statistics and machine learning communities provide collaboration and contribute to development of algorithms in this area.

APPLICATIONS

Logistic regression may be used to calculate the probability of future events based on historical data. This is a fundamental capability for predictive analytics.

A logistic regression model is essentially a calculator that is able to estimate probabilities resulting from combinations and interactions of multiple independent variables. These models are complementary to distribution models discussed previously.

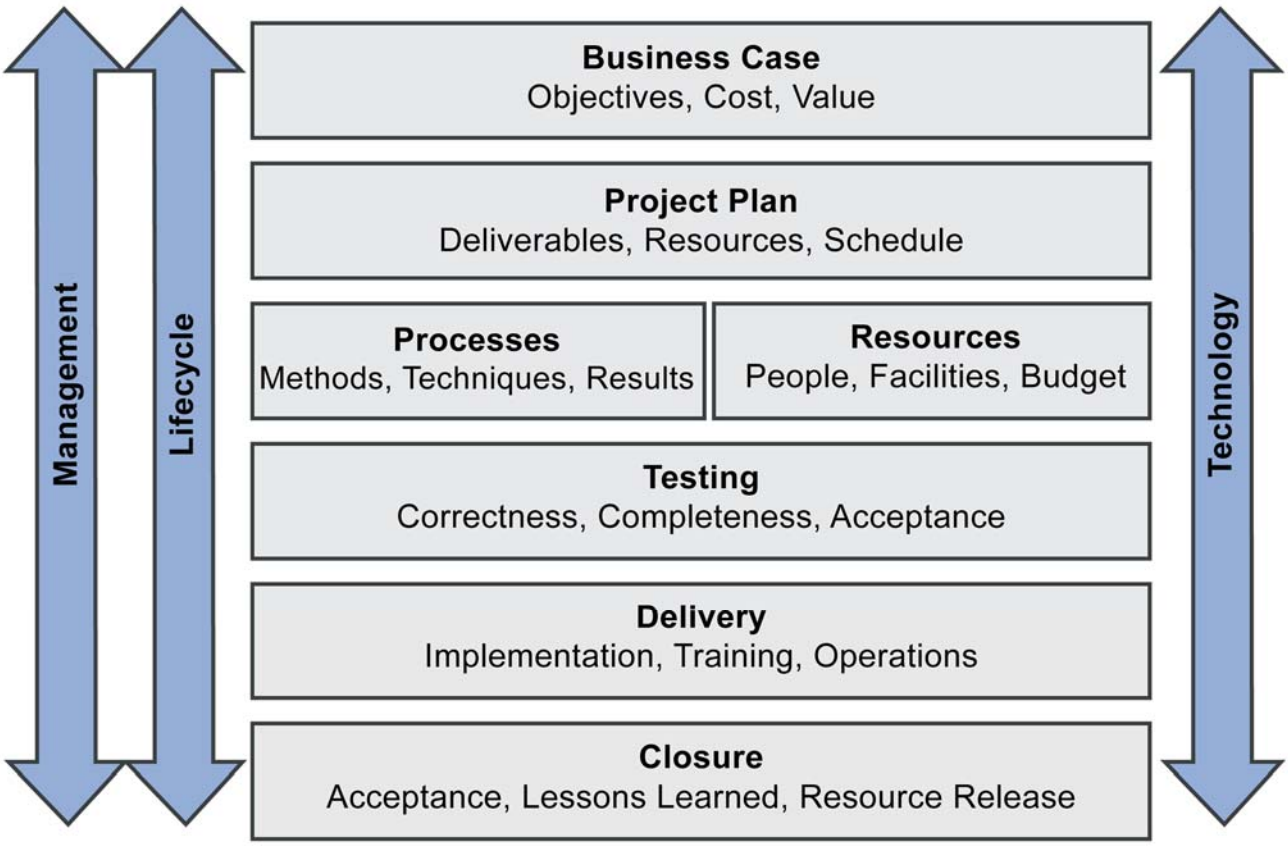
tdwi
Transforming Data With Intelligence™

Module 4

Building Predictive Models

Model Building Processes

Data Mining Projects



Model Building Processes

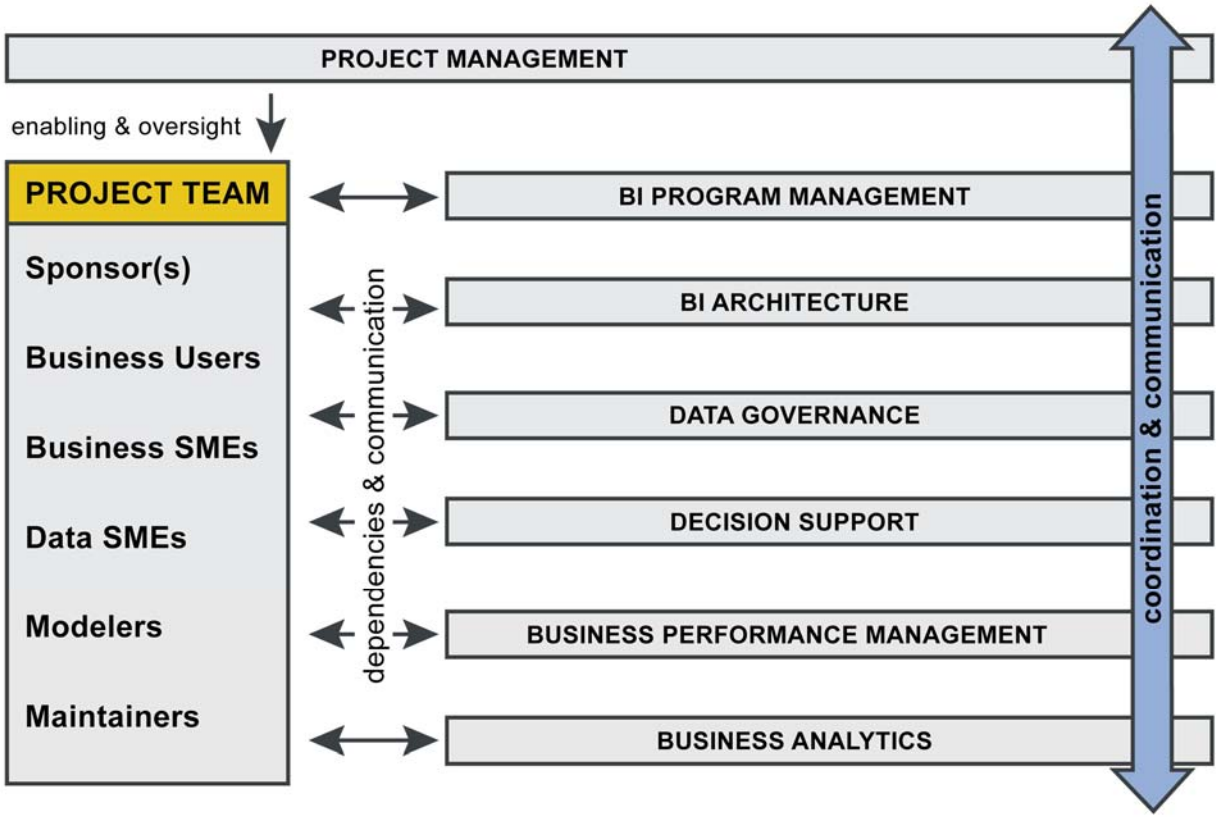
Data Mining Projects

PROJECT DISCIPLINE

Data mining is complex and should be undertaken with the discipline of projects. Avoid the temptation to "jump into the deep end of the data pool," and start with business objectives and benefits. Plan the project before committing resources and performing data mining processes and activities. Test results and then deliver. Finally close the project with formal acceptance of deliverables and a project review. Execute data mining projects with the right level of project management, a defined lifecycle, and the right technology to do the job well.

Implementation and Operations Teams

A Team Effort



Implementation and Operations Teams

A Team Effort

PROJECT TEAM

Data mining is a team effort that involves people with a variety of knowledge and skills. Common team roles include:

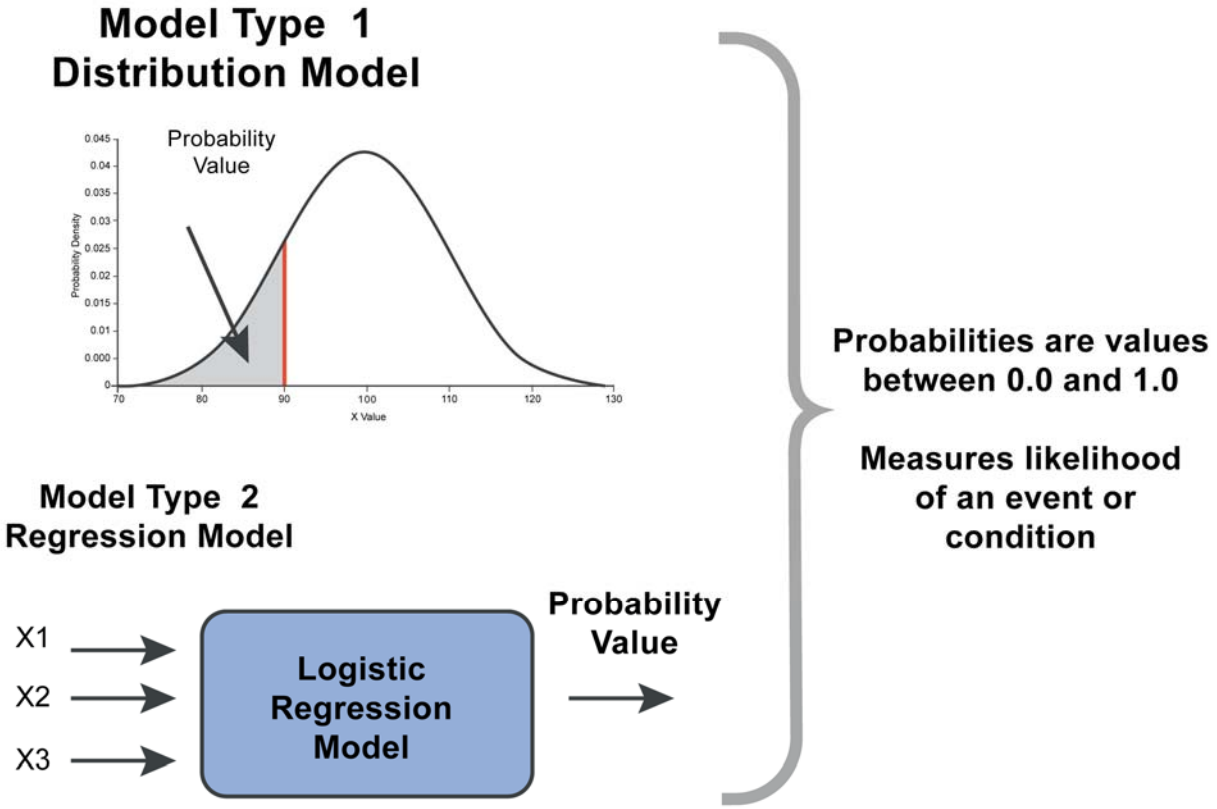
- Sponsorship—securing funding, resources, and political will that are necessary for successful projects
- Business users—real people who will use the results and who understand the business needs
- Business subject experts—providing depth of knowledge about business needs and the business domain
- Data subject experts—providing depth of knowledge about data availability, data sourcing, and data content
- Modelers—with knowledge of applied statistics, data mining techniques, and use of data mining technology
- Maintainers—mining and technically skilled people who will support and maintain models after deployment

PROJECT DEPENDENCIES

Data mining projects are rarely standalone activities. They have dependencies and communication responsibilities with many enterprise functions, organizations, and initiatives—project management, program management, governance—as illustrated on the facing page.

Predictive Techniques

Probability Values



Predictive Techniques

Probability Values

FOUNDATION OF PREDICTION

As described in an earlier section of the course, a probability value is a measure of the likelihood of a future event of condition. It provides the most fundamental form of a prediction.

Probabilities are commonly made at the granularity of an event. Other forms of predictions described in the module consider other forms of granularity of the prediction.

Common modeling approaches for creating probability estimates include distribution models and logistic regression models.

Technology

Features and Functions Overview

Function	Features
data structure	flat file, relational, columnar ...
data storage	file system, in database, in memory ...
data connectivity	ODBC, DBMS gateway, Hadoop ...
metadata	integrated, data source generated ...
model types	descriptive, decision, predictive ...
mining techniques	classify, segment, associate, sequence, predict ...
algorithm types	decision tree, regression, neural net ...
data visualization	graphs, maps, tables, pivots ...
model building	GUI, wizards, workflow, collaboration ...
data preparation	cleansing, transformation, sampling, ...
administration	security, versioning, deployment, monitoring
packaged mining applications	response, risk, activation, attrition, upsell ...
packaged rule models	credit, fraud, profitability ...

Technology

Features and Functions Overview

WHAT DATA MINING TOOLS CAN DO

Data mining tools are more than model building tools. While model building is central to data mining, the tools must support many different functions that are prerequisite, post-requisite, and adjacent to modeling. From data access and preparation to model administration, mature and robust tools do much more than provide algorithms and parameter settings. The facing page illustrates the range of functions and features that are typical of modern data mining technology.

Model Building Algorithms

What and Why

WHAT?

> "A data mining algorithm is a well-defined procedure that takes data as input and produces models or patterns as output."
>
> Sargur Srihari, SUNY Buffalo

WHY?

- ✔ Structure and terminology conventions for data mining methods
- ✔ Precise encoding of a procedure as a finite set of rules
- ✔ Reusable in many business contexts to produce needs-specific models
- ✔ Encapsulates details of statistics and mathematics
- ✔ Reduces error opportunity in model building

COMPONENTS

- ✔ Task—the purpose: classification, clustering, regression, etc.
- ✔ Structure—form of the model: linear regression, hierarchical clustering, etc.
- ✔ Scoring—function to judge model quality: accuracy, error, loss, etc.
- ✔ Searching—method of finding patterns in data: linear, gradient, etc.

Model Building Algorithms

What and Why

DATA MINING ALGORITHM DEFINED

Sargur Srihari, a machine learning educator and researcher at SUNY Buffalo, defines a data mining algorithm as "a well-defined procedure that takes data as input and produces models or patterns as output."

An algorithm by Srihari's definition must:

- Conform to (or occasionally establish) structure and terminology conventions for data mining methods
- Precisely encode a procedure as a finite set of rules—the procedure terminates after a finite number of steps and produces an output
- Be useful in many different contexts to produce models tailored to varying requirements and data
- Encapsulate the details of statistics and mathematical functions in a way that reduces probability of errors when building models

An algorithm includes a task or purpose, structure, a scoring function, and pattern search methods.

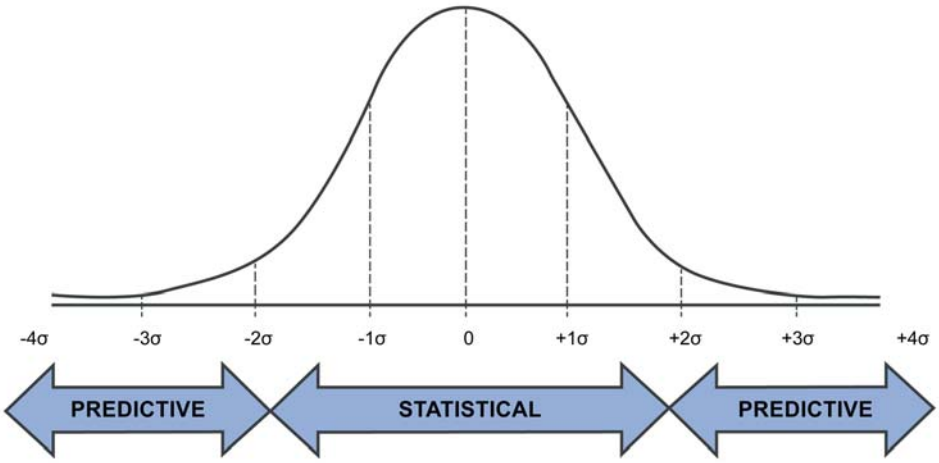
tdwi
Transforming Data With Intelligence™

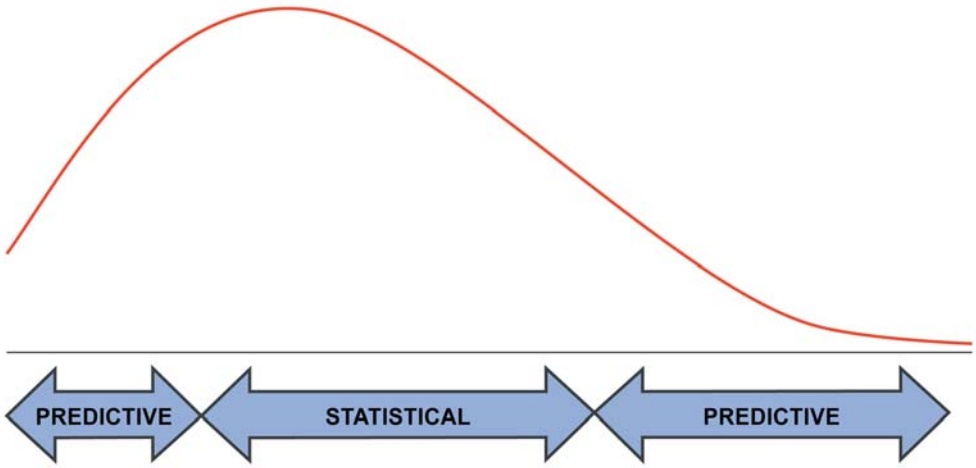
Module 5

Implementing Predictive Capabilities

Introductory Concepts

Distribution View





Introductory Concepts

Distribution View

STATISTICS AND DISTRIBUTION

Traditional statistical analysis is primarily focused on central tendencies—the center of the distribution curve. As variation and standard deviation increase, the information and analytics value declines. This works because the analysis is centered on understanding the nature of outcomes.

PREDICTIVE WITH NORMAL DISTRIBUTION

Predictive analytics shifts the attention away from central tendencies to look at the tails of the curve and things that are distant from central tendencies. In predictive analytics, the purpose is not to understand the nature of outcomes, but to shape future outcomes. Opportunities to enhance business performance are found in the low-incidence, high-impact occurrences in the tails of the distribution. To enhance business performance we must look outside the norm.

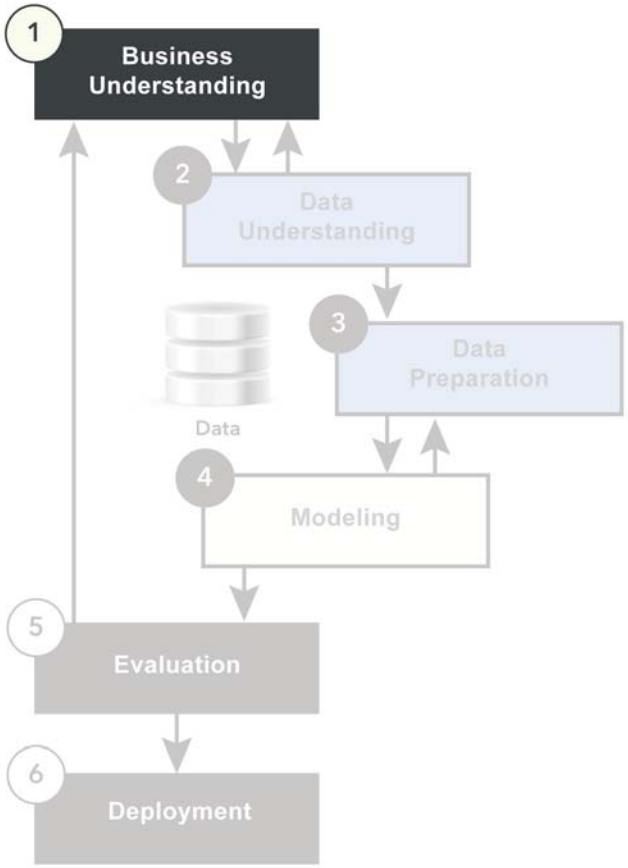
PREDICTIVE WITH SKEWED DISTRIBUTION

Although normal distribution is an important and central concept to analytics, business is not distributed normally in the real world. In predictive analytics (in fact, in all analytics) we must work with skewed distributions.

With the skewed distribution, both tails are still the focus. The longer tail, however, may be the most rewarding. As a practical matter, it is more difficult for predictive modeling to succeed in the tail closest to the mode due to proximity. When successful, it often yields lower impact than a long tail. The reduction in individual behavior impact, however, is partially offset by higher frequency of occurrences in this tail.

Business Understanding

Activities and Deliverables



1.1 Determine Business Objectives
- **1.1.1 Background**
- **1.1.2 Business Objectives**
- **1.1.3 Business Success Criteria**

1.2 Assess Situation
- **1.2.1 Resources Inventory**
- **1.2.2 Requirements, Assumptions, Constraints**
- **1.2.3 Terminology**
- **1.2.4 Risks & Contingencies**
- **1.2.5 Costs & Benefits**

1.3 Determine Data Mining Goals
- **1.3.1 Data Mining Goals**
- **1.3.2 Data Mining Success Criteria**

1.4 Produce Project Plan
- **1.4.1 Project Plan**
- **1.4.2 Initial Tools & Techniques Assessment**

Business Understanding

Activities and Deliverables

BUSINESS OBJECTIVES

Start the process with business and customer perspective. Understanding what the sponsor or customer wants to accomplish is essential to avoid the risk of building elegant models with little or no value. Deliverables from this task include:

Background	Any information about the business circumstances that may affect or be useful to inform the project.
Business Objectives	A description of what the customer wants to accomplish including decisions to be made and questions needing answers.
Business Success Criteria	The measures that will be used to judge usefulness of project results. For subjective criteria, know how and by whom the criteria will be judged.

ASSESS SITUATION

Identify resources, constraints, assumptions, and other factors that influence the analysis goals and that must be considered to develop an objective and realistic project plan. Deliverables include:

Resources Inventory	A list of people, data, tools, and technology available for the project.
Requirements, Assumptions, and Constraints	Project requirements including schedule, quality, and security; assumptions about business and data; resource and technology constraints.
Terminology	A glossary of important language for the project including both business terms and data mining terms.
Risks & Contingencies	A list of factors that might cause project delay or failure, along with mitigation and contingency plans.
Costs & Benefits	An assessment of cost-effectiveness of the project weighing anticipated value against estimated costs.

DETERMINE DATA MINING GOALS

Extend from business objectives (in business terms) to describe the project goals in technical terms. Deliverables include:

Data Mining Goals	A description of project outputs and their relationships to the business objectives.
Data Mining Success Criteria	Technical success criteria for the project such as information quality and predictive accuracy.

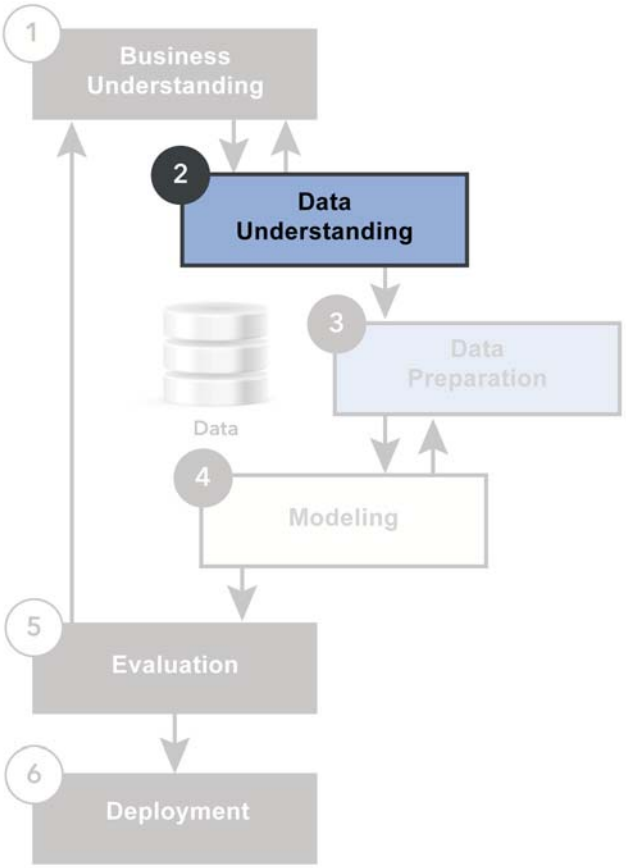
PRODUCE PROJECT PLAN

Develop a project plan that accounts for iteration and includes initial view of mining tools and techniques. Deliverables include:

Project Plan	Typical phase-by-phase plan and schedule for the project.
Tools & Techniques Assessment	Initial concept of data mining techniques to be used and the choice of tools to apply those techniques.

Data Understanding

Activities and Deliverables



2.1 Collect Initial Data
2.1.1 Initial Data Collection Report

2.2 Describe Data
2.2.1 Data Description Report

2.3 Explore Data
2.3.1 Data Exploration Report

2.4 Verify Data Quality
2.4.1 Data Quality Report

Data Understanding

Activities and Deliverables

COLLECT INITIAL DATA

Gain access and acquire the data identified as project resources. Load the data into any tools that will be used for data exploration and quality assessment. If using data from multiple sources, begin to think about data integration here. The deliverable is:

Initial Data Collection Report	Create a record of the data set(s) acquired, their locations, and the methods used to acquire them. Document any problems that occurred and the methods of resolving problems.

DESCRIBE THE DATA

Take an "outside looking in" approach to describe the externally visible properties of acquired data prior to more detailed data exploration. The deliverable is:

Data Description Report	Report properties such as data format, quantities (number of tables, rows, columns, etc.), key fields and identifiers if known, and other "surface" features. Assess the degree to which the data satisfies project needs.

EXPLORE THE DATA

Look beyond externally visible data characteristics to understand the shape and characteristics of the data using statistical and visualization techniques. The deliverable is:

Data Exploration Report	Report all of the basic statistics for attributes that are important to the project including mean, median, mode, maximum value, minimum value, and sums for some simple aggregations. Report and visualize frequency distribution of key attributes and describe characteristics such as percent of null values and percent unique.

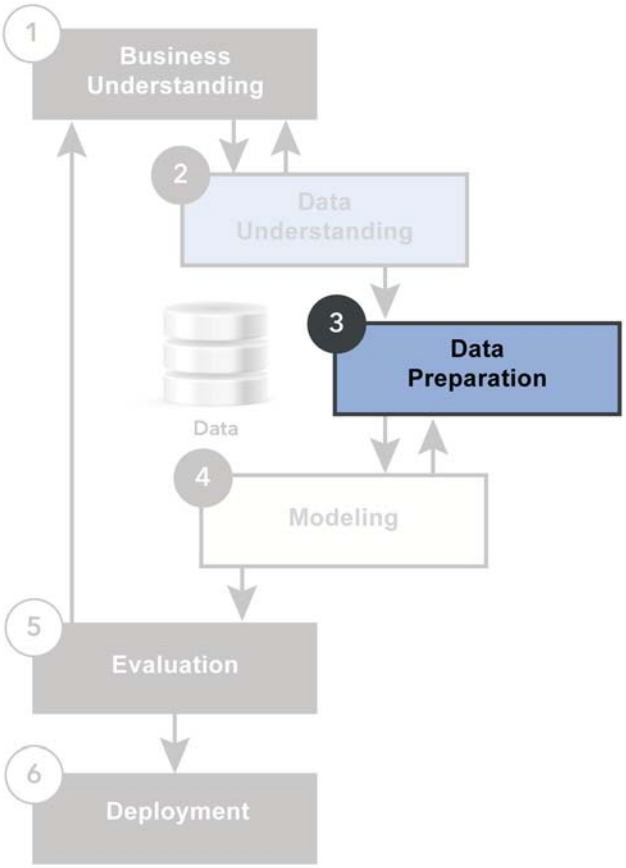
VERIFY DATA QUALITY

Examine data quality with attention to quality criteria of completeness, accuracy, and consistency. Explore questions such as what the most common errors are and where they occur. The deliverable is:

Data Quality Report	Describe the results of data quality examination including errors found, frequency with which they occur, impact on data mining goals, and possible solutions when needed.

Data Preparation

Activities and Deliverables



3.1 Select Data
- **3.1.1 Rationale for Inclusion/ Exclusion**

3.2 Clean Data
- **3.2.1 Data Cleaning Report**

3.3 Construct Data
- **3.3.1 Derived Attributes**
- **3.3.2 Generated Records**

3.4 Integrate Data
- **3.4.1 Merged Data**

3.5 Format Data
- **3.5.1 Reformatted Data**

Data Preparation

Activities and Deliverables

SELECT DATA

Decide which data to use for analysis. Consider criteria of relevance to data mining goals, quality, and technical constraints such as limits on data volume or data types. Specify selection for attributes (columns) and for records (rows) from data sources. The deliverable is:

Data Selection Rationale	List the data to be included and the data to be excluded. Describe the reasons for the choices.

CLEAN DATA

Improve data quality as needed by applying data cleansing techniques such as filtering, default values, estimation of most probable values, and so on. The deliverable is:

Data Cleansing Report	Describe what decisions and actions were taken to improve data quality.

CONSTRUCT DATA

Build the data set to be used in modeling including derived attributes, generated records, and data transformations. The deliverables are:

Derived Attributes	New attributes that are constructed from one or more existing attributes in the same record. Example: area = length * width.
Generated Records	Create entirely new records when needed for modeling, For example, create records for customers who made no purchase during the past year.

INTEGRATE DATA

Combine data from multiple sources and prepared data sets to create new records or values. The deliverable is:

Merged Data	Consolidated data from multiple disparate data sources.

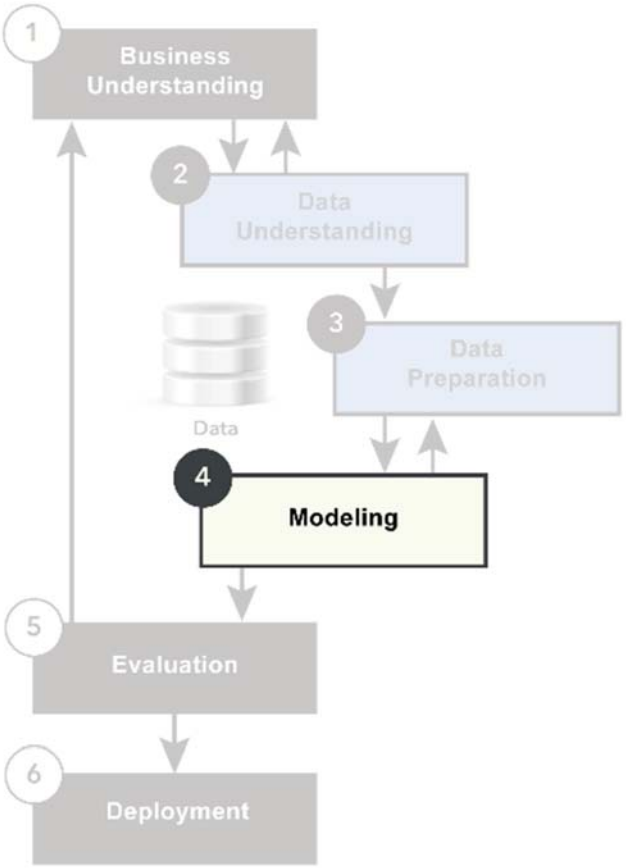
FORMAT DATA

Apply data transformations needed for syntactic and representational consistency across the entire modeling data set. The deliverable is:

Reformatted Data	A modeling-ready data set that accounts for tool requirements such as attribute types and formats, record sequence, attribute sequence, record identifiers, etc.

Modeling

Activities and Deliverables



4.1 Select Modeling Technique
- **4.1.1 Modeling Technique**
- **4.1.2 Modeling Assumptions**

4.2 Generate Test Design
- **4.2.1 Test Design**

4.3 Build Model
- **4.3.1 Parameter Settings**
- **4.3.2 Models**
- **4.3.3 Model Description**

4.4 Assess Model
- **4.4.1 Model Assessment**
- **4.4.2 Revised Parameter Settings**

Modeling

Activities and Deliverables

SELECT MODELING TECHNIQUE

Identify the modeling technique to be used. If multiple techniques are applied, perform this task separately for each technique. The deliverables are:

Modeling Technique	Document the selected modeling technique with rationale.
Modeling Assumptions	Identify and document assumptions driven by the selected technique. Many modeling techniques make specific assumptions about the data—for example, that all attributes have uniform distributions, etc.

GENERATE TEST DESIGN

Determine how you will test the model before building the model. What are the quality and validity criteria, and how will you test that they are satisfied? What basis will you use to separate the data into training and testing sets? The deliverable is:

Test Design	Describe the plan to train, test, and evaluate the models. Give special attention to the way training and testing data will be separated.

BUILD MODEL

Run the modeling tool with training data to create one or more models. The deliverables are:

Parameter Settings	Settings and rationale for each parameter specified by the technique and tool.
Models	The actual models generated by the modeling tool.
Model Description	Model documentation and interpretation.

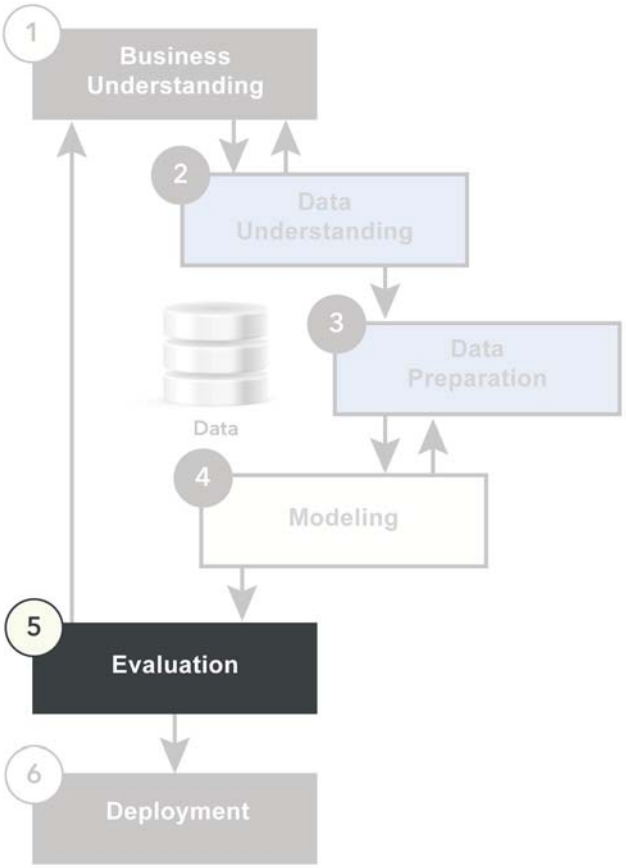
ASSESS MODEL

Evaluate the degree to which the models meet data mining goals and satisfy data mining success criteria. The deliverables are:

Model Assessment	Summarize results of the assessment and rate the quality of multiple models relative to each other.
Revised Parameter Settings	Based on indications from assessment, fine tune the models by adjusting parameter settings in preparation for another cycle of model building activity.

Evaluation

Activities and Deliverables



5.1 Evaluate Results
- **5.1.1 Assessment of Modeling Results**
- **5.1.2 Approved Models**

5.2 Review Process
- **5.2.1 Review of Process**

5.3 Determine Next Steps
- **5.3.1 List of Possible Actions**
- **5.3.2 Decision**

Evaluation

Activities and Deliverables

EVALUATE RESULTS

The assessment activities of modeling address model evaluation from an "inside the model" point of view. This phase evaluates the ability of the model to meet business objectives, satisfy business needs, and measure up to business success criteria. The deliverables are:

Assessment of Data Mining Results	An assessment that closes the loop back to business success, evaluating how results match business success criteria defined at the start of the project.
Approved Models	Identification of models that meet business success criteria and that are approved for deployment.

REVIEW PROCESS

Review the processes used by the project, identifying activities that were missed, performed poorly, and performed well. Include quality of process and quality of solution as part of the review. The deliverable is:

Review of Process	Summarize the process review to highlight activities that have been missed and those that should be repeated.

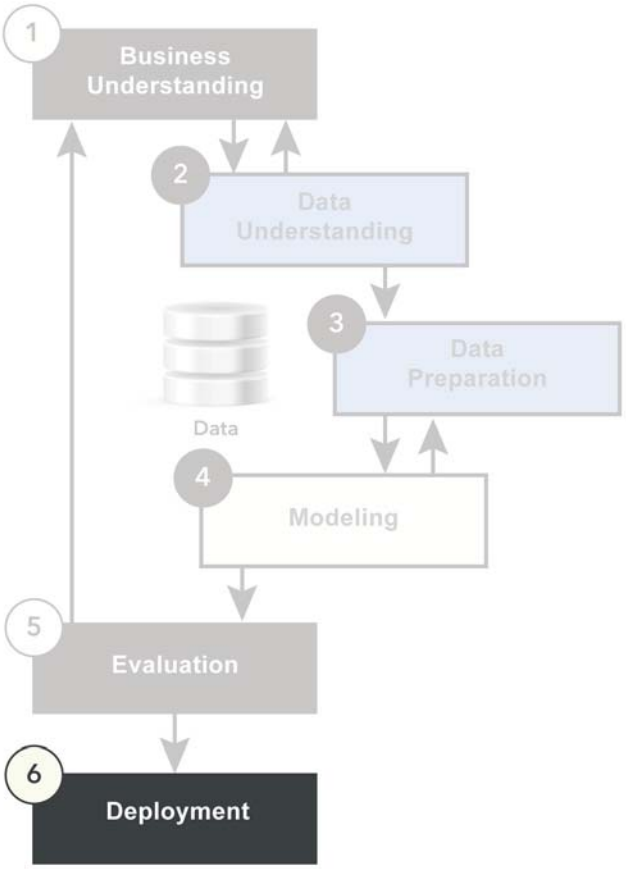
DETERMINE NEXT STEPS

Use the results of assessment and process review to decide whether to deploy the models, adjust and iterate, or initiate new projects. Consider quality of results, meeting of business objectives, and state of schedule and budget in these decisions. The deliverables are:

List of Possible Actions	List the possible actions with pros and cons of each.
Decision	Describe decisions, rationale, and expected results.

Deployment

Activities and Deliverables



6.1 Plan Deployment
- **6.1.1 Deployment Plan**

6.2 Plan Monitoring & Maintenance
- **6.2.1 Monitoring & Maintenance Plan**

6.3 Produce Final Report
- **6.3.1 Final Report**
- **6.3.2 Final Presentation**

6.4 Review Project
- **6.4.1 Experience Documentation**

Deployment

Activities and Deliverables

PLAN DEPLOYMENT

Working from assessment results, develop a plan to deploy the models. Create, apply, or evolve repeatable processes for model deployment as much as is practical. The deliverable is:

Deployment Plan	Summarize the deployment plan, identifying steps, sequence, and guidelines.

PLAN MONITORING AND MAINTENANCE

When data mining results are intended as part of the day-to-day business processes, monitoring and maintenance must be considered. Develop a plan to avoid misunderstanding and misuse of data mining results. The deliverable is:

Monitoring and Maintenance Plan	Summarize the monitoring and maintenance plan, identifying steps, sequence, and guidelines.

FINAL REPORT

As part of project closure, produce a final report of project results. Depending on sponsor and business expectations, the report may range from a final status report to an in-depth presentation for business stakeholders. The deliverables are:

Final Report	A final written report meeting stakeholder needs and expectations.
Final Presentation	A formal presentation of the final report when needed.

PROJECT REVIEW

Review and document experiences from which you can learn and improve as well as the very positive experiences that you want to be able to repeat. The deliverable is:

Experience Documentation	A summary of project learning, recommendations for improvement in future projects, and recommendations for repeatable processes and practices for future projects.

Module 6

Human Factors in Predictive Analytics

Analytics Culture

Executive Buy-In

Analytics Culture

Executive Buy-In

CULTURE DEFINED

In business, culture relates to the shared values, attitudes, standards, and beliefs that characterize the people in an organization and define the nature of that organization. Organizational culture is rooted in goals, strategies, structure, and relationships with employees, customers, stakeholders, and community. Culture is an essential component in success or failure of many business endeavors including business analytics.

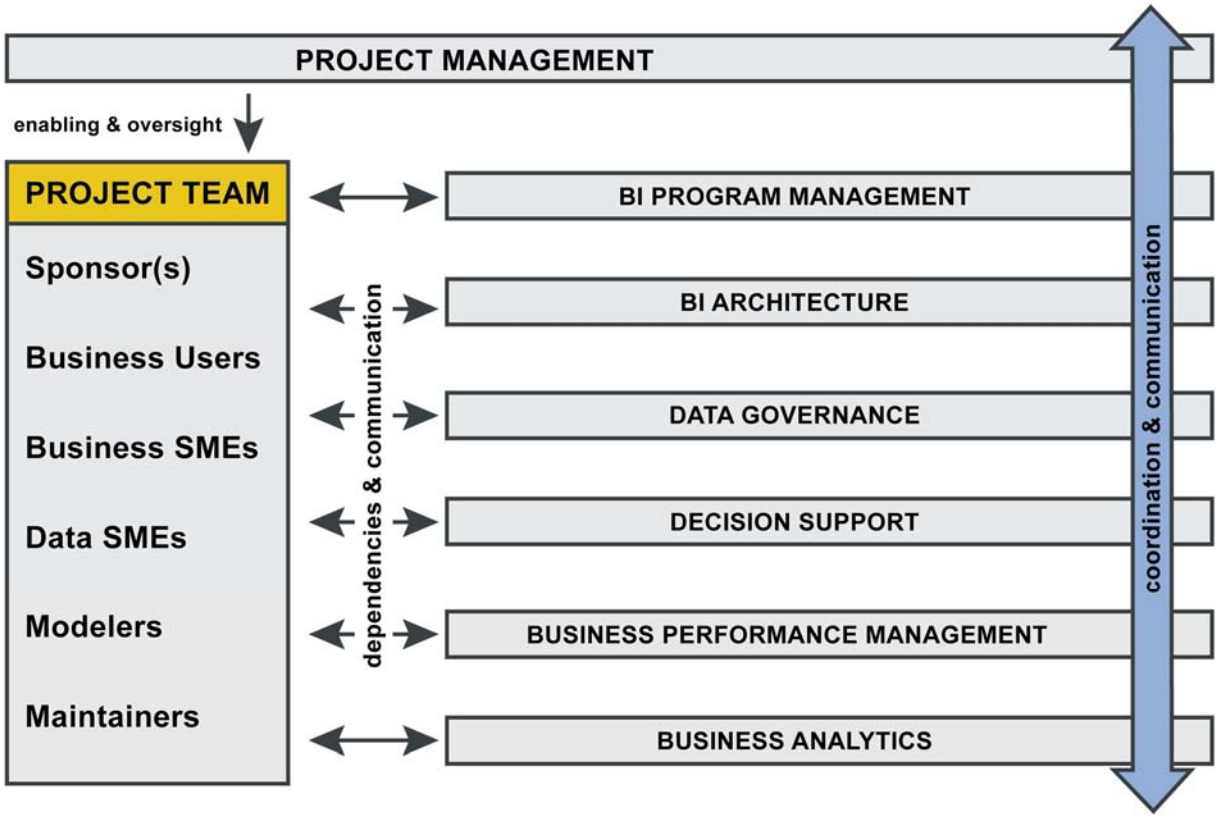
SHAPING ANALYTICS CULTURE

"Culture always starts with the owner."[1] This quote captures the essence of analytics sponsorship and the importance of executive buy-in. When the executives demonstrate trust in analytics, that trust tends to permeate through the layers of the organization. When they visibly and vocally support analytics as a core decision-making and management competency, then others follow with open support. When resources for analytics are provided at the top, then middle and line managers tend to find budget, people, and time to dedicate to their local analytics needs.

[1] Patrick, Josh [2013]. "The Real Meaning of Corporate Culture," *New York Times*, May 21. http://boss.blogs.nytimes.com/2013/05/21/the-real-meaning-of-corporate-culture/?_r=0)

People and Predictive Analytics

The Team



People and Predictive Analytics

The Team

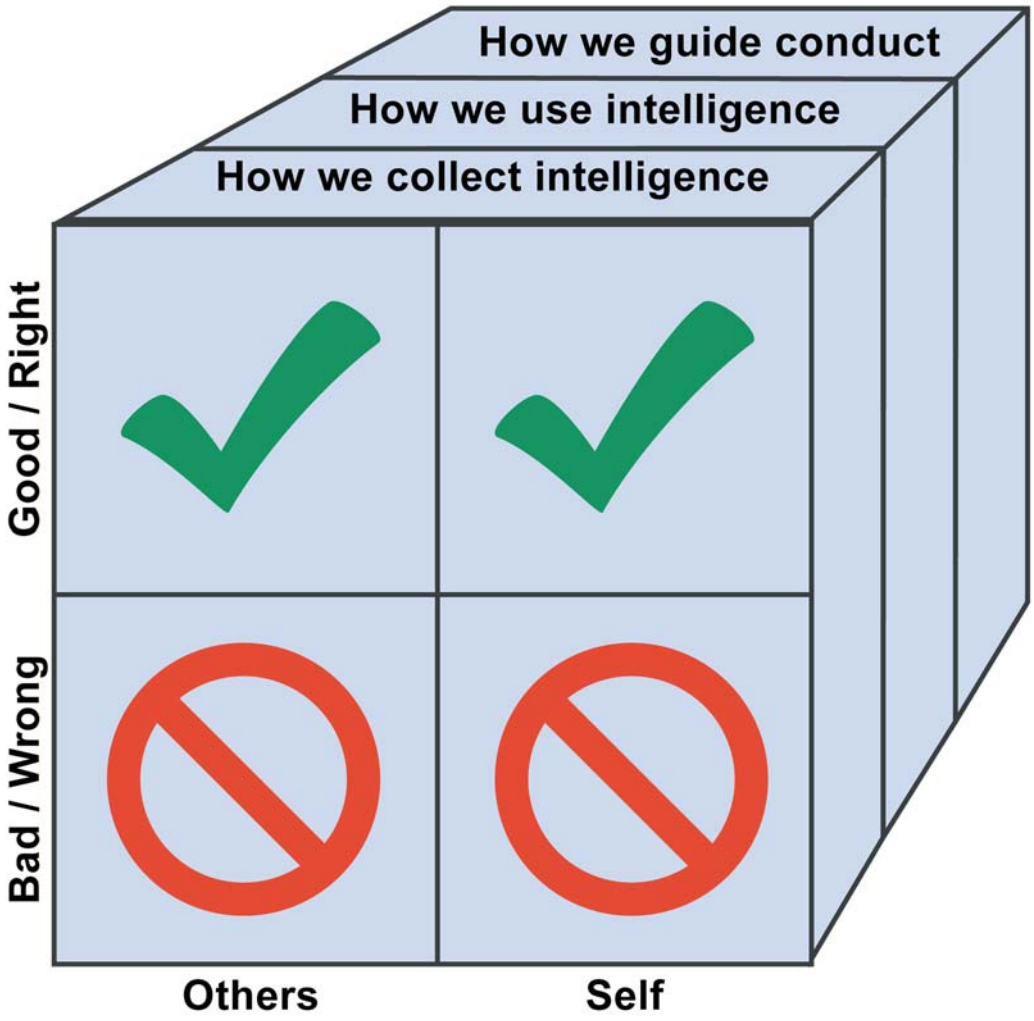
PROJECT TEAM

Recall that we previously discussed predictive analytics as a team effort including roles of:

- Sponsorship—securing funding, resources, and political will that are necessary for successful projects
- Business users—real people who will use the results and who understand the business needs
- Business subject experts—providing depth of knowledge about business needs and the business domain
- Data subject experts—providing depth of knowledge about data availability, data sourcing, and data content
- Modelers—with knowledge of applied statistics, data mining techniques, and use of data mining technology
- Maintainers—mining and technically skilled people who will support and maintain models after deployment

Ethics and Predictive Analytics

Why Ethics Matters



Ethics and Predictive Analytics

Why Ethics Matters

WHAT IS ETHICS

Ethics is the challenge of distinguishing right from wrong and good from bad. It is challenging because "right" and "good" are not always clear. Furthermore, right and wrong must account for all stakeholders. It is common to think of ethics as doing right for others, but when doing right for others is harmful to the self, ethics is not served. Nor is ethics served when doing right for self is harmful to others. Ethics has two aspects—the tension of right vs. wrong and the tension of right vs. right.

ETHICS, BI, AND ANALYTICS

Business intelligence raises many new ethical questions about how we collect and use intelligence, and how those choices guide and shape organizational and individual conduct. Every decision that we make and every action that we take shapes the perceptions of customers, employees, partners, competitors, and the public.

We are approaching an era where every BI program will need to actively manage ethics. More data, more kinds of data, and advanced analysis of data often conflict with concerns of data privacy, security, anonymity, and ownership. Especially in the area of predictive analytics and predictions of individual behaviors, ethics is significant.

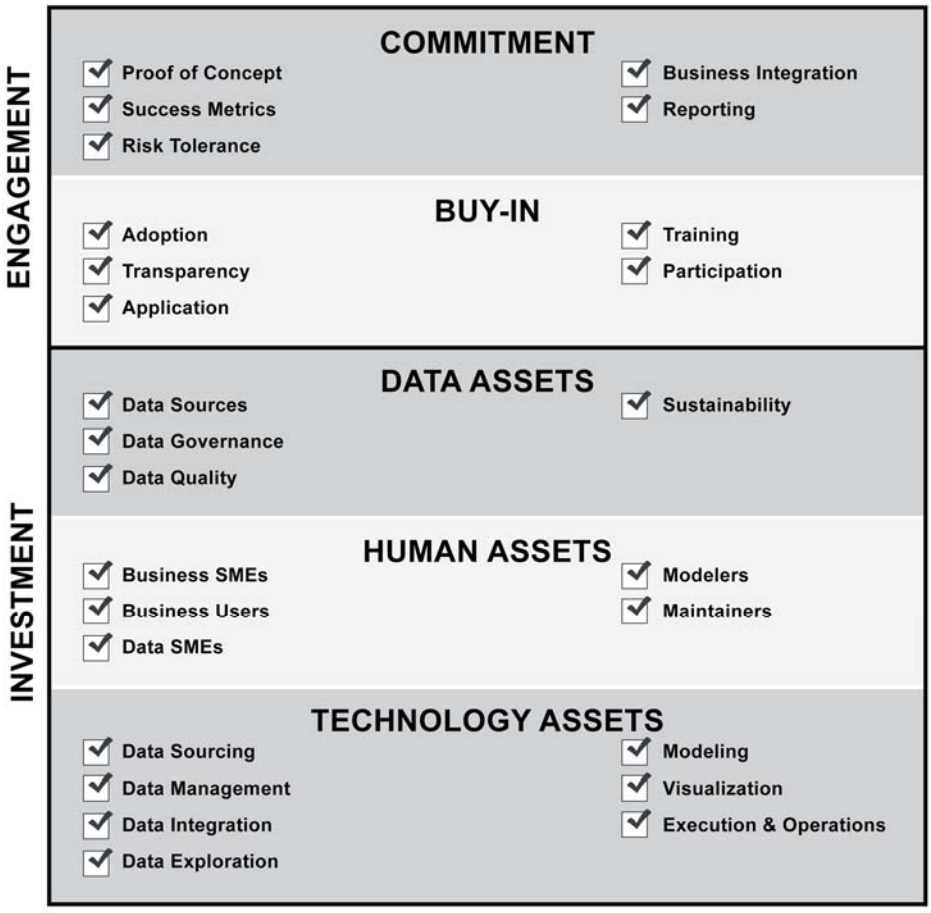
Transforming Data With Intelligence™

Module 7

Getting Started with Predictive Analytics

Predictive Analytics Readiness

Readiness Checklist



Predictive Analytics Readiness

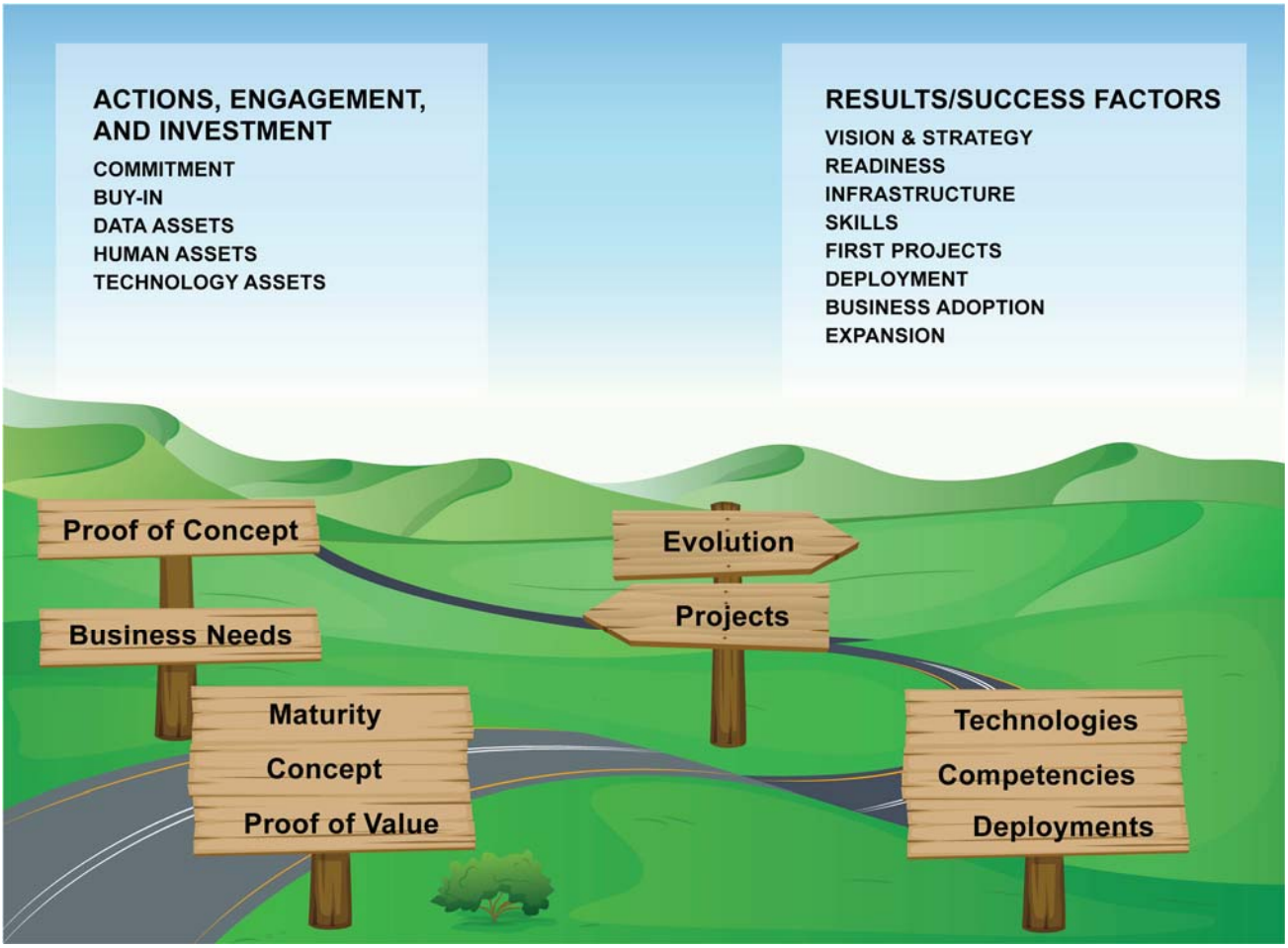
Readiness Checklist

ASSESSING THE CURRENT STATE

Think of predictive analytics as a journey along the path of maturing organizational intelligence and decision capabilities. As with any journey it is important to understand your current position before taking the first step. The readiness checklist on the facing page describes several categories to consider when evaluating your current position. Getting started in the right way is primarily about determination and the assets that you have to enable success—engagement in the form of commitment and buy-in that is supported with data, human, and technology assets.

Predictive Analytics Roadmap

A Plan to Evolve



Predictive Analytics Roadmap

A Plan to Evolve

VISION AND STRATEGY

A roadmap is a plan that matches short-term and long-term goals with steps and activities designed to help meet those goals in an organized way. The roadmap has three purposes:

- Building consensus about needs and the people, processes, solutions, and technologies required to satisfy those needs
- Looking into the future to understand dependencies and anticipate sequence and timing of projects, technologies, and results that are needed to satisfy the goals
- Providing a framework for more detailed planning of projects and assets to accomplish the short-term and long-term goals

A predictive analytics roadmap ideally accounts for all of the readiness factors from commitment to technology assets, planning to grow capabilities and evolve ever-increasing readiness through a sequence of:

- Articulating the vision and strategy
- Assessing initial readiness
- Getting the infrastructure in place
- Acquiring the essential skills to get started
- Executing and learning from first projects
- Deploying business solutions
- Growing business adoption
- Expanding solutions, adoption, and analytics maturity

This page intentionally left blank.