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.

TDWI Predictive Analytics Fundamentals

The Data Warehousing Institute 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:

email: info@tdwi.org

Publication Date: October 2013

TABLE OF CONTENTS

COURSE OBJECTIVES

To learn:

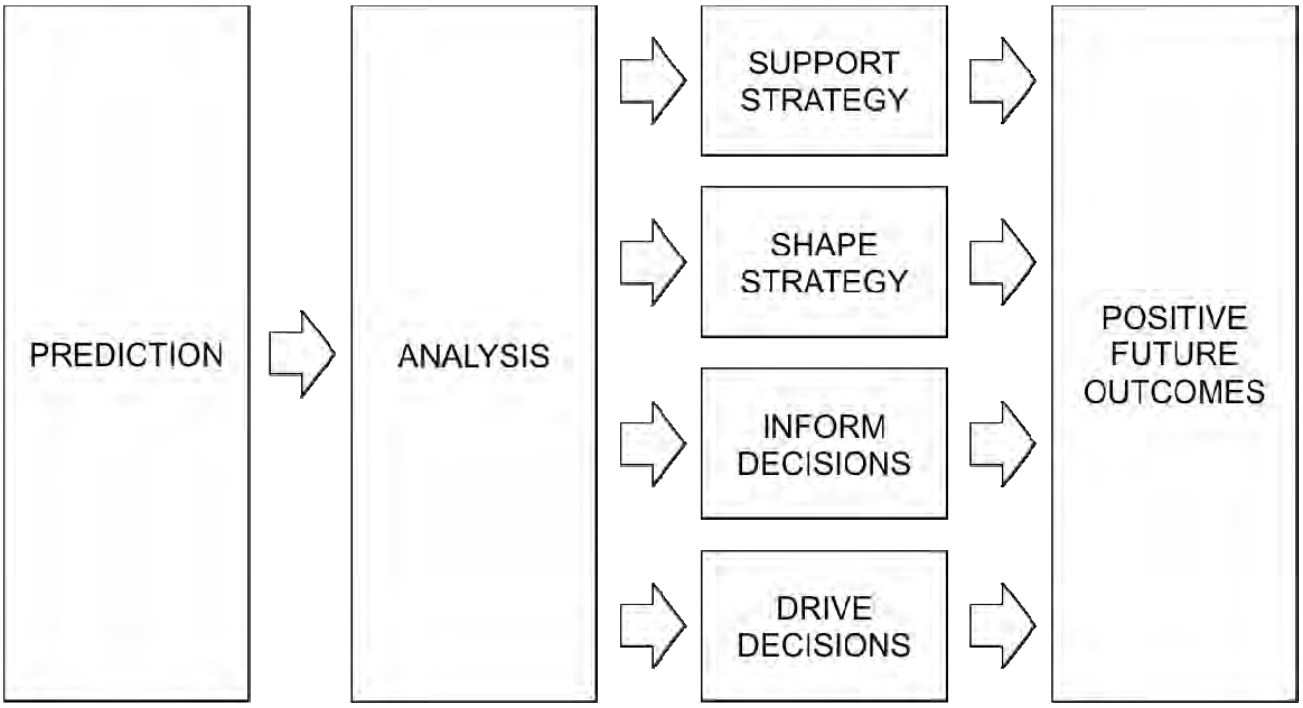
- ✓ ***Definitions, concepts, and terminology of predictive analytics***
- ✓ ***Common applications of predictive analytics***
- ✓ ***How and where predictive analytics fits into a BI program and the relationships with business metrics, performance management, and data mining***
- ✓ ***To distinguish among various predictive model types and understand the statistical foundations of each***
- ✓ ***Organizational considerations for predictive analytics including roles, responsibilities, and the need for business, technical, and management skills***
- ✓ ***Practical guidance for getting started with predictive analytics***

Module 1

Predictive Analytics Concepts

What and Why of Predictive Analytics

Business Value of Predictive Analytics



What and Why of Predictive Analytics

Business Value of Predictive Analytics

SHAPING YOUR BUSINESS FUTURE

The purpose of predictive analytics is to provide insights that inform business processes in which business activities affect future business outcomes. Prediction provides information about the probability of future behaviors – the probability, for example, that a specific customer group has a particularly high risk of moving to a competitor. Analysis adds meaning to the prediction with context and depth – understanding why the customer segment is an attrition risk. The combination of prediction and understanding is applied to achieve one or more of:

- Supporting strategy – When a core element of business strategy is an exceptionally high rate of customer retention, then identification of at risk groups is essential.

- Shaping strategy – Assuming that the at risk group has a common demographic profile (young technology professionals, for example) then adapting strategy to improve appeal to that demographic segment can help to achive business objectives.

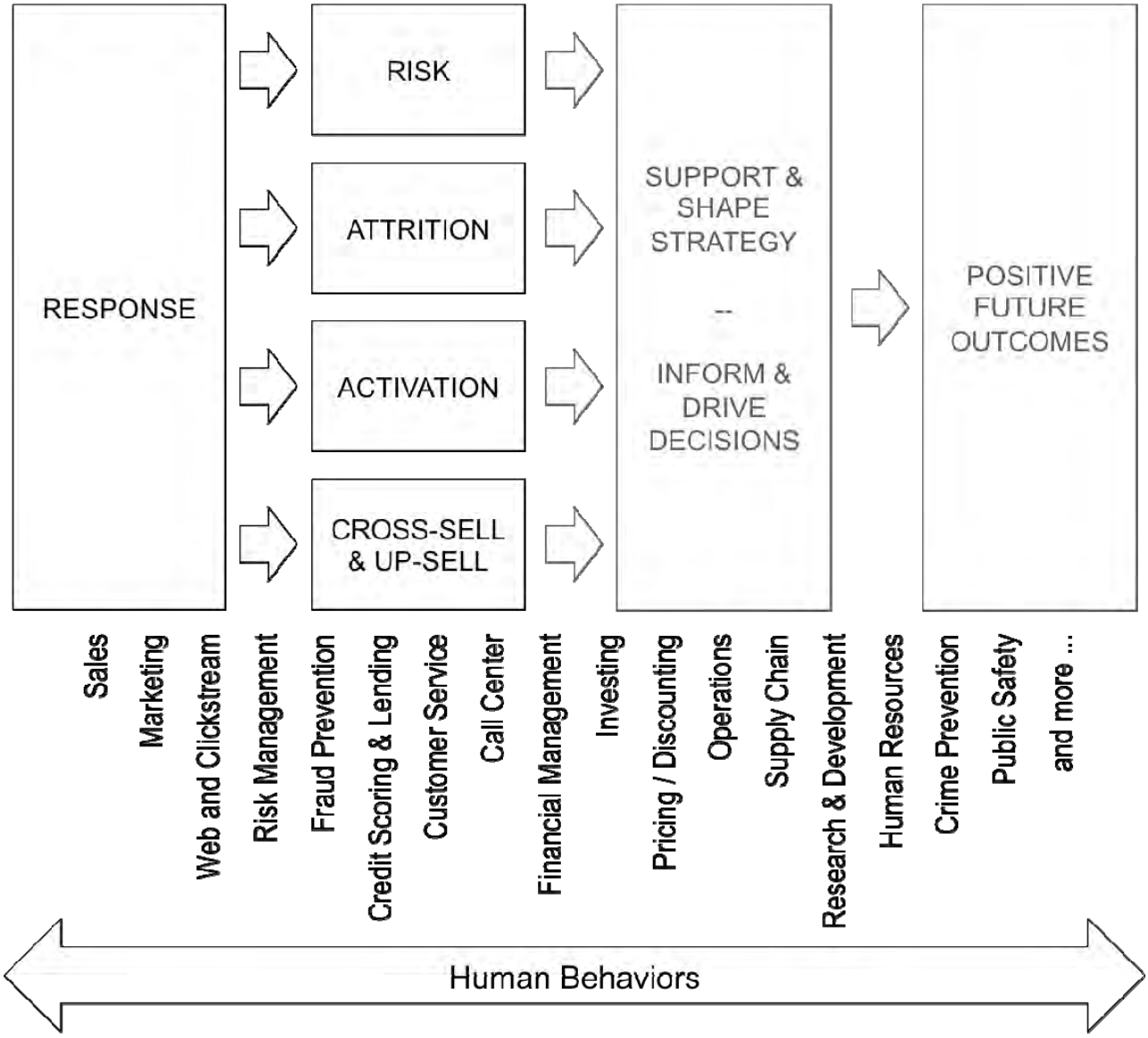
- Inform decisions – With knowledge of attrition risk probabilities, tactical decision makers can make informed decisions to allocate resources to manage the risk, such as dedicating staffing and financial resources for outreach to the at risk group.

- Drive decisions – With reliable models and the right business rules in place, predictive analytics may trigger automated offers of discounts and other benefits for contract extension by customers in at risk groups.

The purpose for each of these uses is to drive positive future business outcomes. Business value is achieved when outcomes support business objectives, applications of predictive analytics drive positive outcomes, analysis guides the application of analytics, and prediction shapes the focus of analysis.

Common Applications for Predictive Analytics

What Business Needs to Predict



Common Applications for Predictive Analytics

What Business Needs 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.

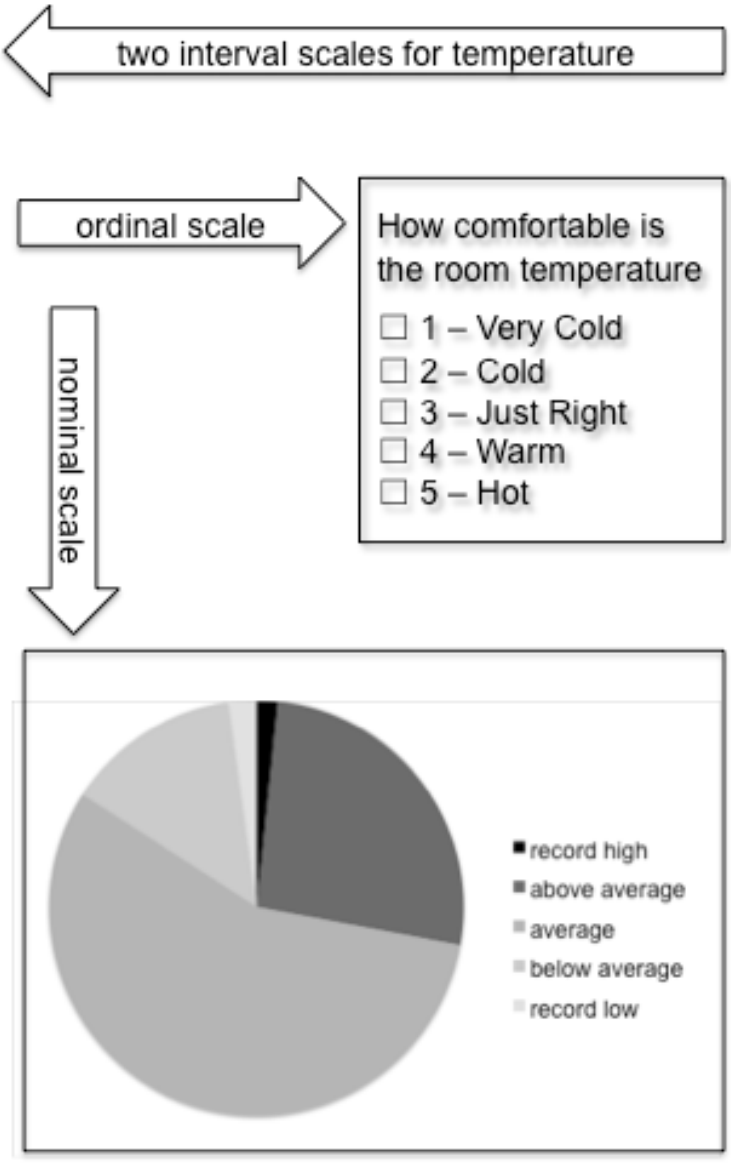
Module 2

Data Mining Fundamentals

Statistics and Data Mining

Variables

Fahrenheit	Celsius
32	0
40	4
50	***10***
60	**16**
70	21
80	27
90	**32**
100	***38***
110	43
120	49
130	54
140	60
150	66
160	71
170	77
180	82
190	88
200	93
210	99
212	100



okay to calculate?	nominal	ordinal	interval	ratio
frequency distribution	YES	YES	YES	YES
median & percentiles	NO	YES	YES	YES
sum & difference	NO	NO	YES	YES
mean & standard deviation	NO	NO	YES	YES
ratio	NO	NO	NO	YES

Statistics and Data Mining

Variables

RATIO, INTERVAL, ORDINAL, AND NOMINAL

Classifying variables as ratio, interval, ordinal and nominal describes the types of measurement scales that can be used and the kinds of statistical and mathematical functions that can be applied.

Ratio scales are perhaps the most recognized of measurement scales. Virtually all-physical measures – mass, length, velocity, etc. – are ratio scales. A ratio scale is distinguished by the fact that is has a non- arbitrary zero value, and that all other values are relative to zero. Comparing ratio variables, then, becomes standardized – twenty is always five times larger than four. All arithmetic operations and all statistical functions can be applied to ratio measurements.

With Interval scales, there is no absolute zero point, which limits the ability to compare values. Units of the scale are equally distributed as with a ratio scale, but the zero point is arbitrary. A common example of an interval scale is temperature – Fahrenheit and Celsius – where zero degrees is arbitrary. Using the Fahrenheit scale, it doesn't make sense to say that 100 degrees is twice as hot as 50 degrees. If that were true, then when we look at the Celsius scale 38 would be twice as large as 10. Ratio arithmetic – multiply and divide – don't apply, but add/subtract analysis is useful. Many common statistical functions (correlation, regression, variance, mean, standard deviation) can be applied.

Ordinal scales place values in rank order. The values are non-proportional but comparative. Unlike interval scales where the units are evenly distributed, units of an ordinal scale may be uneven. Using a five-star rating system, for example, we know that a five star hotel is better than a four star hotel, and that a four star is better than three. But we don't know how much better. The level of improvement from three to four may be moderate while the improvement from four to five is substantial. With ordinal measures simple add and subtract operations are meaningless. Median and mode are the only practical statistical functions for ordinal variables.

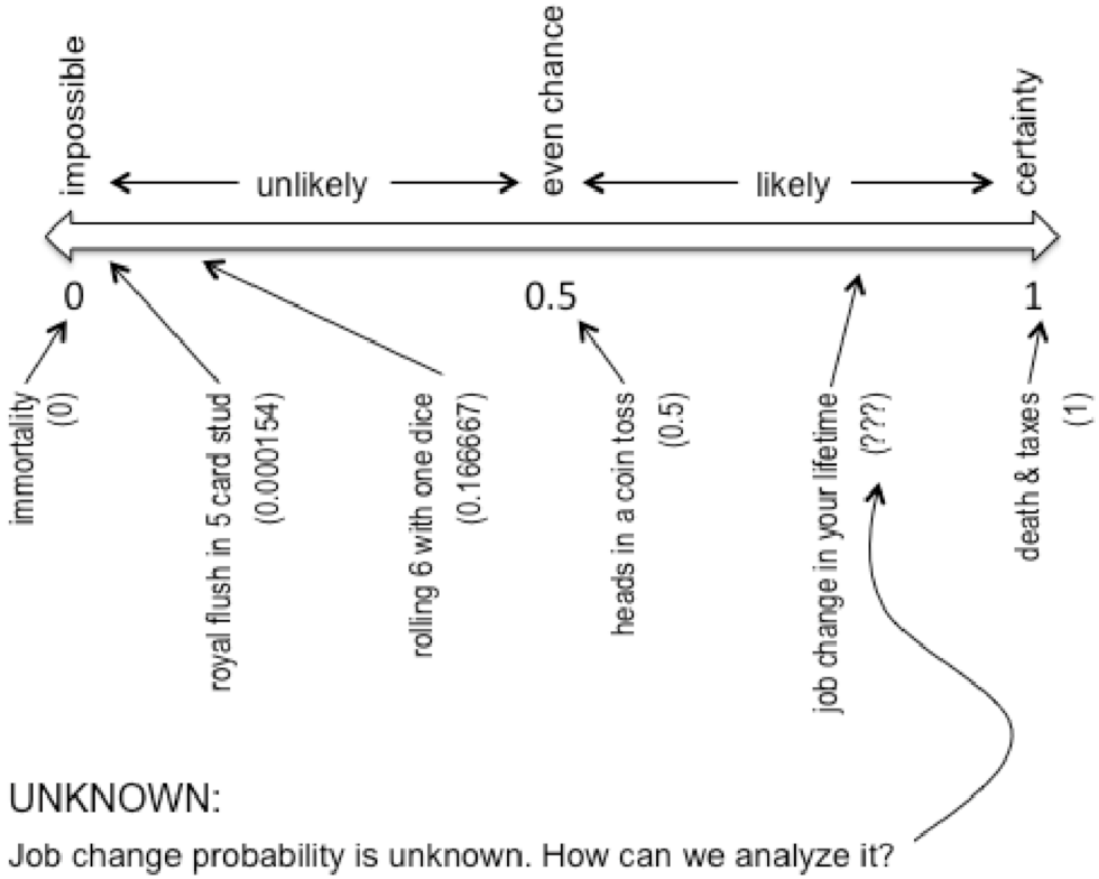
Nominal scales are really an implementation of set theory as a means to measure a population. Members are placed into named sets based on attributes. Variables assessed on a nominal scale are known as categorical variables – they are used to categorize. Arithmetic functions cannot be applied for a nominal scale, and mode is the only statistical function that applies.

Statistics and Data Mining

Probability

PROBABILITY:

A measure of how likely it is that something will occur. The value of a probability measure is always in the range of zero to one.



UNKNOWN:

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

DATA MINING:

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?

Statistics and Data Mining

Probability

MEASURING PROBABILITY

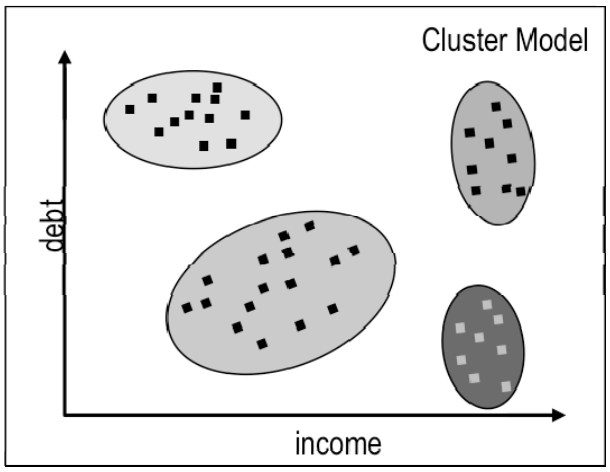
Probability is an important concept in predictive analytics. FICO, the decision management company best known for credit scoring, says that "predictive analytics turn uncertainty into usable probability." In statistical terms, probability measures how likely it is that something will occur. The value 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 at the top of the facing page illustrates several examples of probability between the two extremes.

ANALYZING PROBABILITY

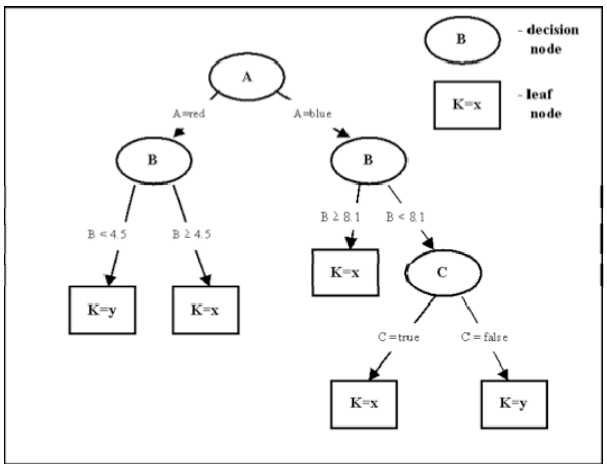
Probability analysis is investigation and study to turn uncertainty into a usable probability measure. Many of the examples shown here are mathematically certain – probability of head in a coin toss is always 0.5 – without need for probability analysis. One of the examples – job change – is uncertain and a good candidate for analysis. Data mining provides the means to perform that analysis beginning with problem context and ending with a useful probability measure.

Data Mining Models

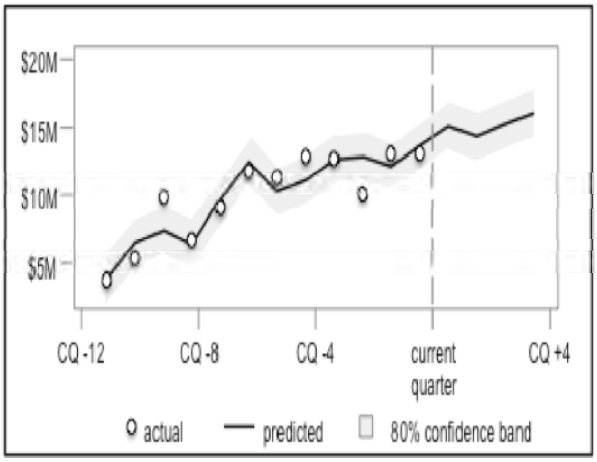
Kinds of Models



Descriptive Model
describe data relationships
use to separate into groups



Decision Model
map influences of decision variables
use to forecast outcomes of actions



Predictive Model
predict behaviors of people
use to drive business results

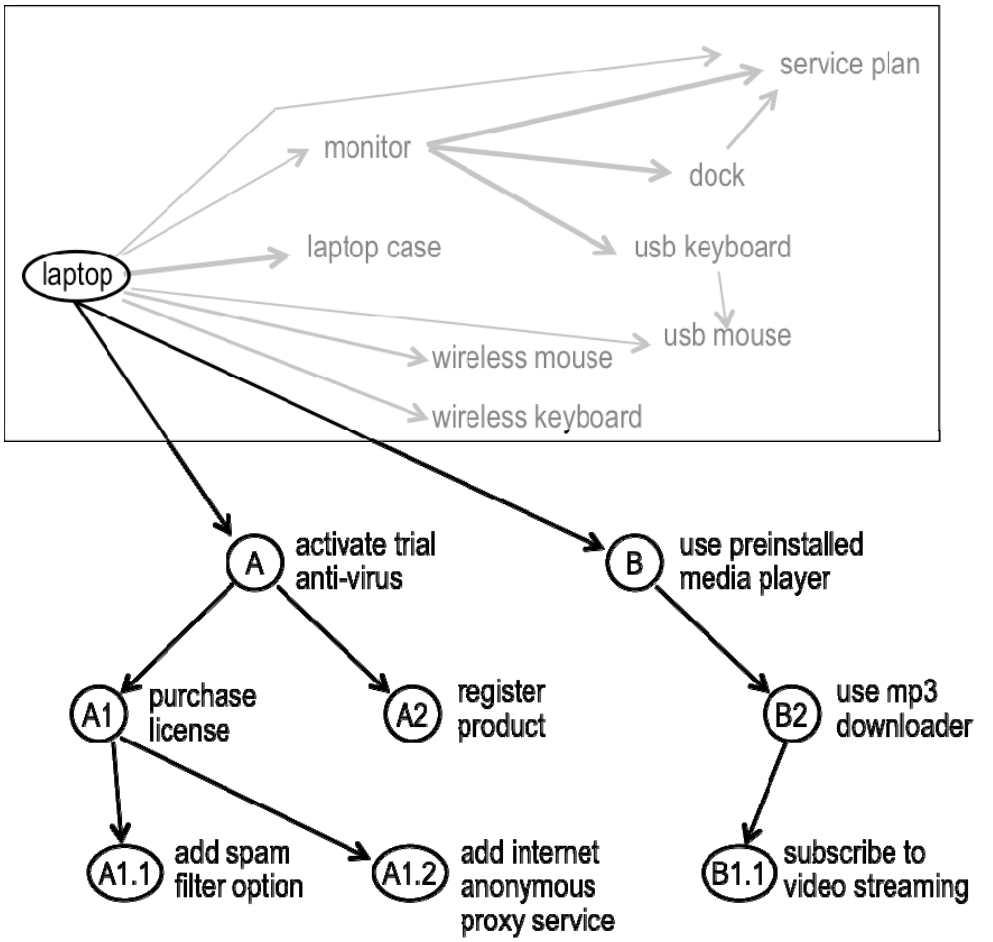
Data Mining Models

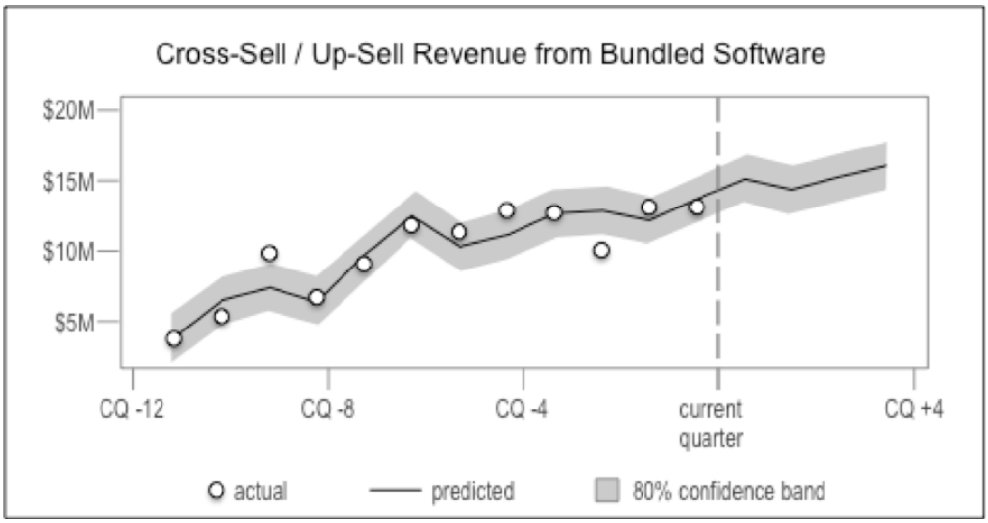
Kinds of Models

DESCRIPTIVE MODELS	Descriptive models describe relationships found in data. They identify relationships between things – customers, products, etc. – in a way that is useful to classify the members of a population into groups.
DECISION MODELS	Decision models forecast the outcomes of complex decisions by mapping the influences among all of the elements of a decision to estimate the expected results of decisions and actions.
PREDICTIVE MODELS	Predictive models analyze past performance to "predict" how likely an individual (or group of similar individuals) is to exhibit a specific behavior in the future – for example, the probability of a customer to defect to a competitor or of a customer to default on a loan.

Data Mining Techniques

Forecasting





Data Mining Techniques

Forecasting

LOOKING INTO THE FUTURE

Forecasting is the use of historic data to understand future trends and outcomes. Forecasting builds upon association and sequencing techniques to look into the future. By applying time-series analysis techniques it becomes practical to analyze data from the past to estimate what is likely to happen in the future.

It is the nature of time-series data that makes forecasting possible. Time-series data, unlike static data, data can be related to itself – same store, same period sales in prior years, for example. There are many different forecasting methods, and data mining tools offer a variety of forecasting algorithms.

FORECASTING VS. PREDICTION

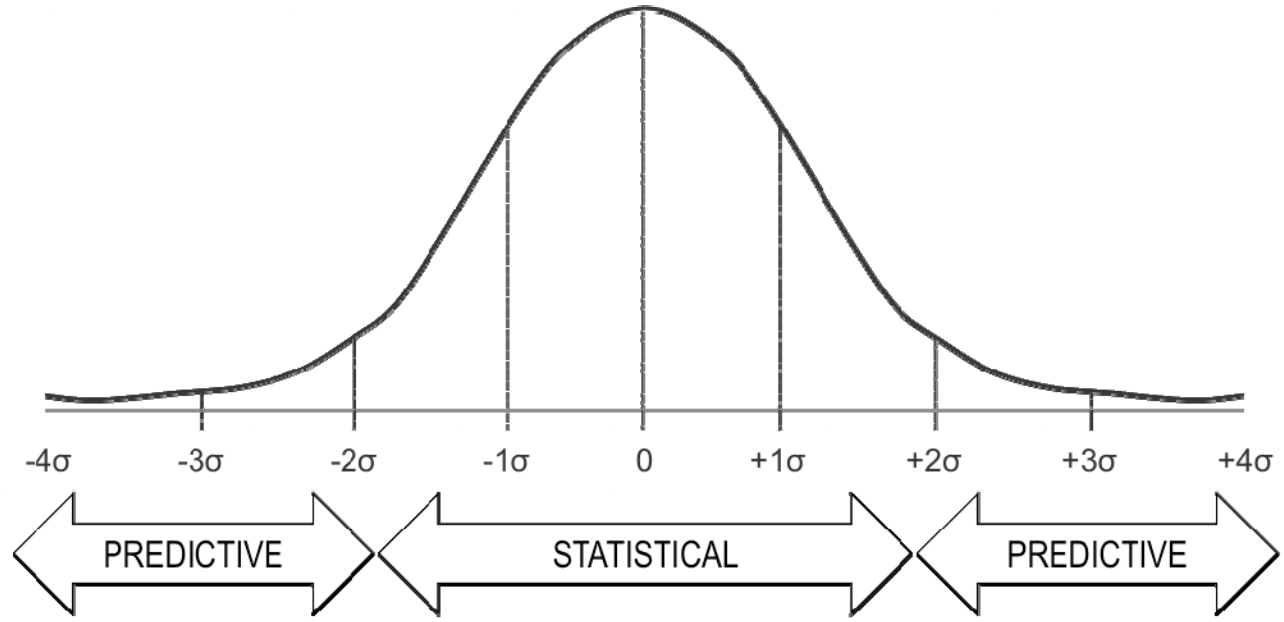
Forecasting and prediction are similar in their purpose of providing information about events that have not yet occurred, they are distinctly different in many other respects. Where forecasting provides aggregate measures – for example, next twelve months customer churn rate by wireless plan – prediction identifies the individual customers (or very fine-grained customer segments) where churn will occur.

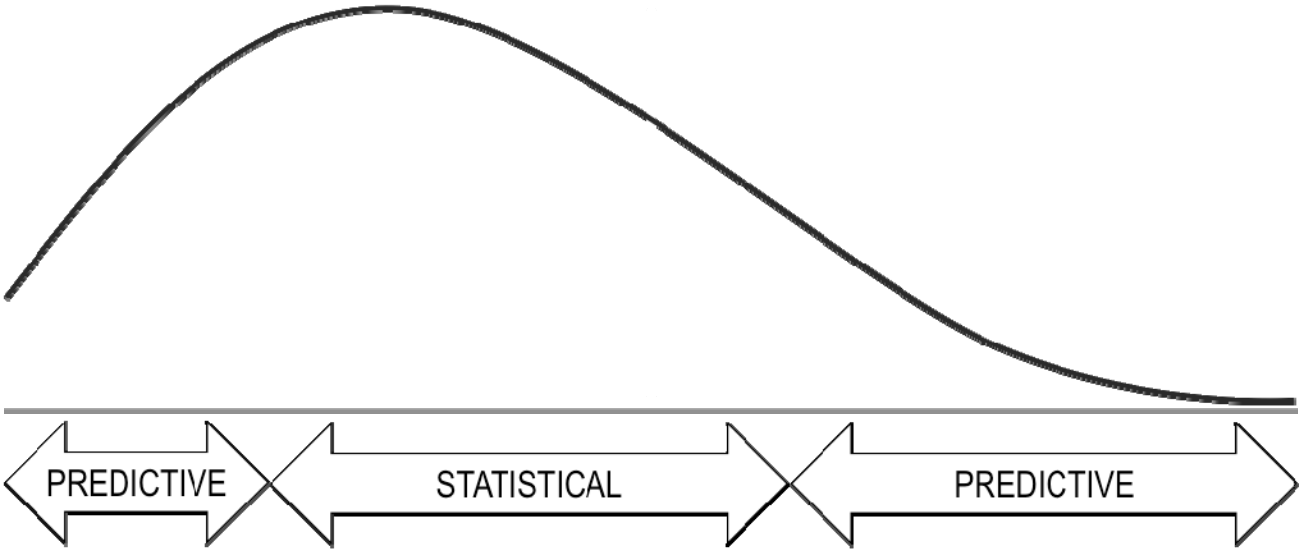
Module 3

Predictive Mining and Modeling

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 analytic value declines. This works because the analysis is centered around 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.

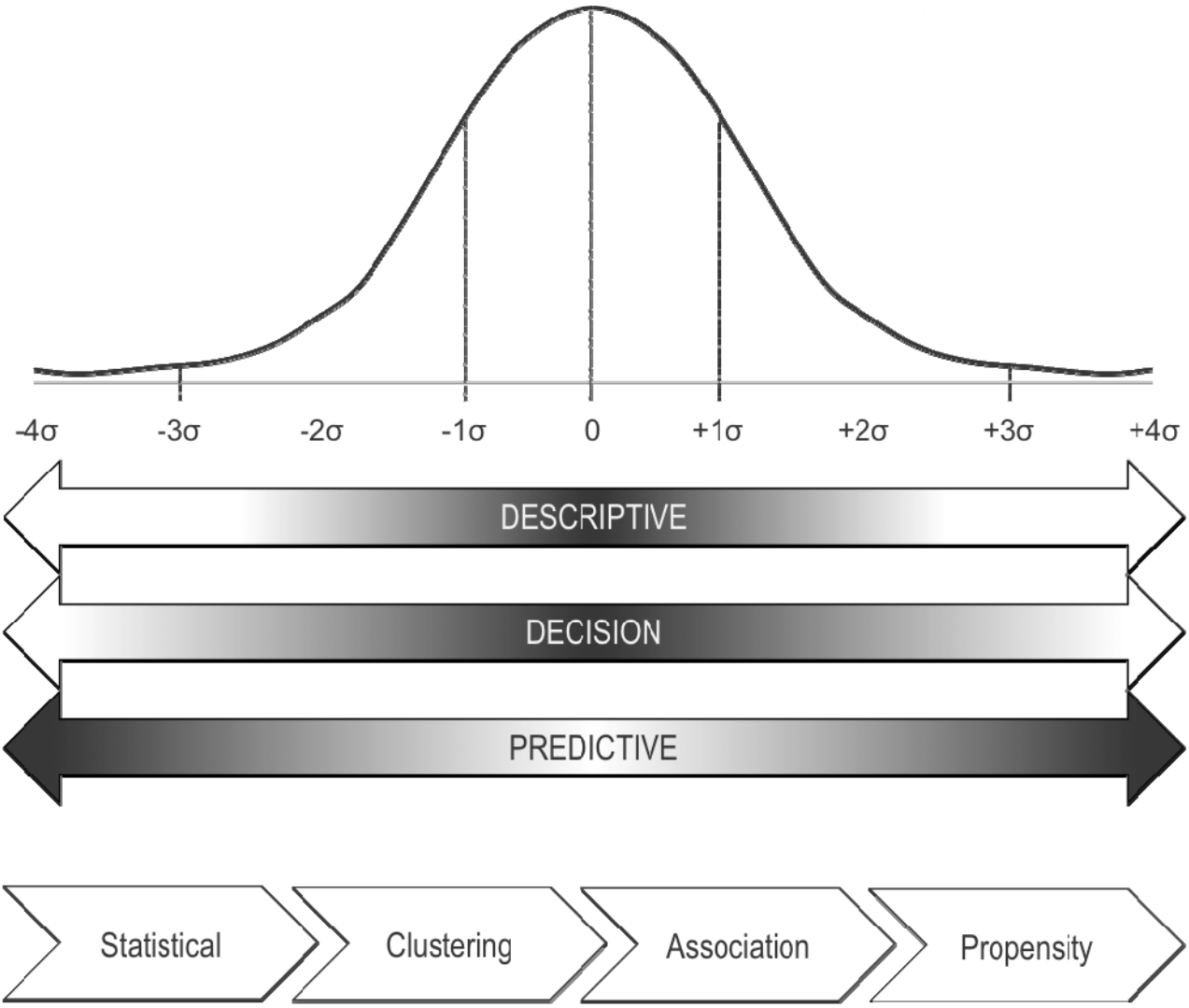
PREDICTIVE WITH SKEWED DISTRIBUTION

While 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 of predictive analytics. 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 in this tail.

Introductory Concepts

Model Types View



Introductory Concepts

Model Types View

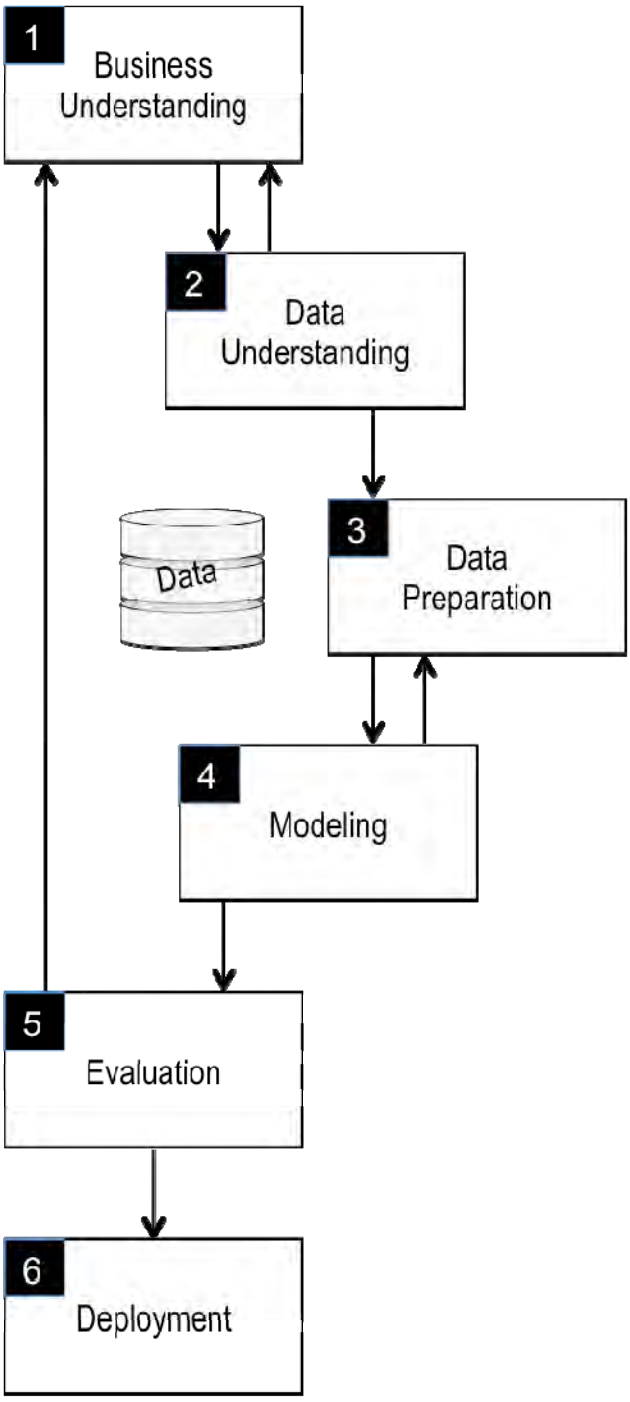
TRADITIONAL COMPLEMENTS PREDICTIVE

Traditional statistical models are not counter to predictive models; they are complementary. Typically you need to understand central tendencies as part of the understanding that is needed to model the distribution tails effectively. Descriptive models help to understand the shape of the data. Decision models help to understand the business purpose and define the objectives for predictive modeling.

Analysis is a process that progresses through steps and stages. There is a natural flow from statistical analysis to overview of the structure and shape of the data, to clustering to find natural groupings, to association for discovery of behavioral relationships among items, and then to propensity to predict future behaviors of individuals.

Introductory Concepts

Process Overview



1.1 Determine Business Objectives
1.2 Assess Situation
1.3 Determine Data Mining Goals
1.4 Produce Project Plan

2.1 Collect Initial Data
2.2 Describe Data
2.3 Explore Data
2.4 Verify Data Quality

3.1 Select Data
3.2 Clean Data
3.3 Construct Data
3.4 Integrate Data
3.5 Format Data

4.1 Select Modeling Technique
4.2 Generate Test Design
4.3 Build Model
4.4 Assess Model

5.1 Evaluate Results
5.2 Review Process
5.3 Determine Next Steps

6.1 Plan Deployment
6.2 Plan Monitoring & Maintenance
6.3 Produce Final Report
6.4 Review Project

Introductory Concepts

Process Overview

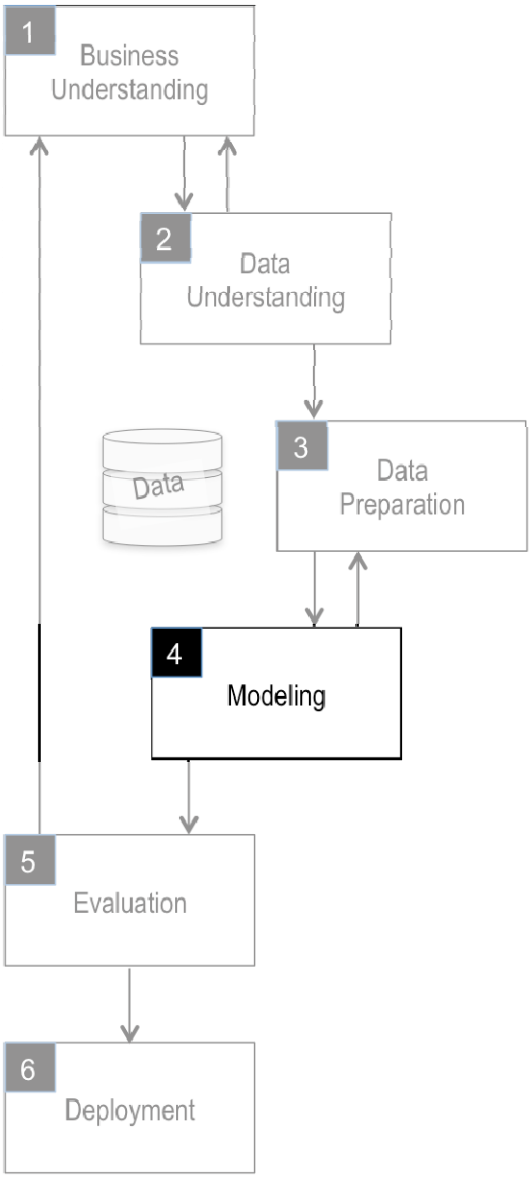
THE TASKS OF CRISP-DM

We've already discussed the phases of CRISP-DM at a high level. At the next level, the methodology defines a set of tasks for each phase. The twenty-four tasks, organized by phase, are illustrated on the facing page.Each task has a defined set of outputs or deliverables that are described through the rest of this module.

The following information is based on the paper *CRISP-DM 1.0 Step-by-Step Data Mining Guide* (http://www.the-modeling-agency.com/crisp-dm.pdf).

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 that is 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, no null values, 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 dataset into trainig and testing sets. The deliverable is:

Test Design	Describe the plan tp train, test, and evaluate the models. Give special attention to the way that 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.

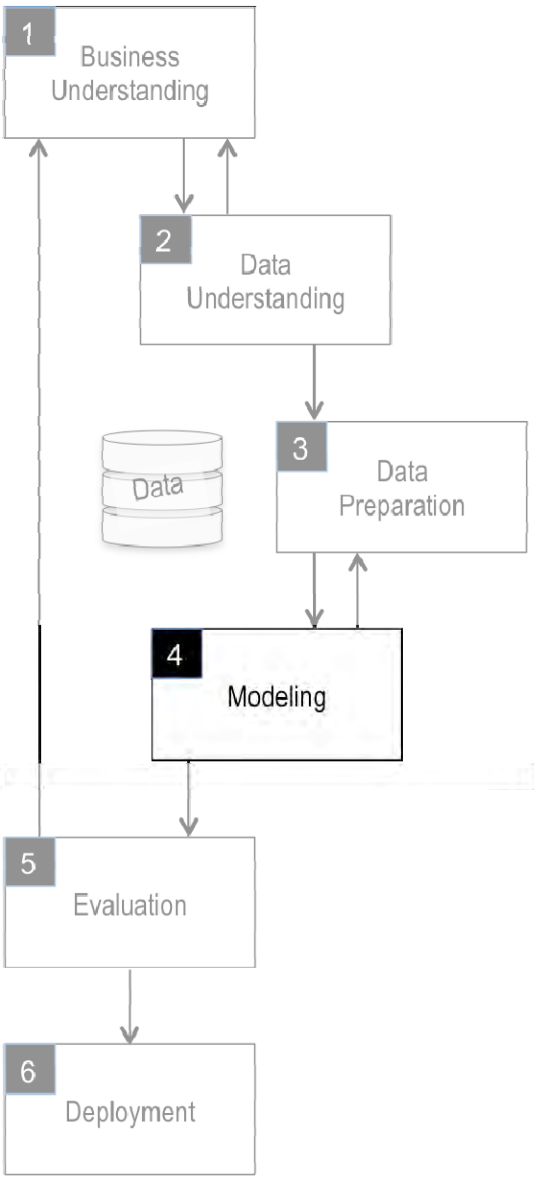
ASSESS MODEL

Evaluate the degree to which the models meet data mining goals and satisft 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 prapration for another cycle of model build activity.

Modeling

Pragmatics



4.1 Select Modeling Technique
4.2 Generate Test Design
4.3 Build Model
4.4 Asses Model

- ✓ Focus on the goals: revisit modeling goals and success criteria when choosing modeling techniques
- ✓ Keep it simple: focus on a single behavior in a single model; when you need to predict many behaviors, work incrementally and expect many models
- ✓ Test before building: test design informs appropriate choice of modeling techniques
- ✓ Separate testing data from training data: divide the data into a training set for machine learning and a testing set for model assessment
- ✓ Randomize test and training data selection: random record selection helps to optimize accuracy without overfitting
- ✓ Test after building: check that the model meets the goals and is a good fit using the test data set

Modeling

Pragmatics

BUILDING USEFUL MODELS

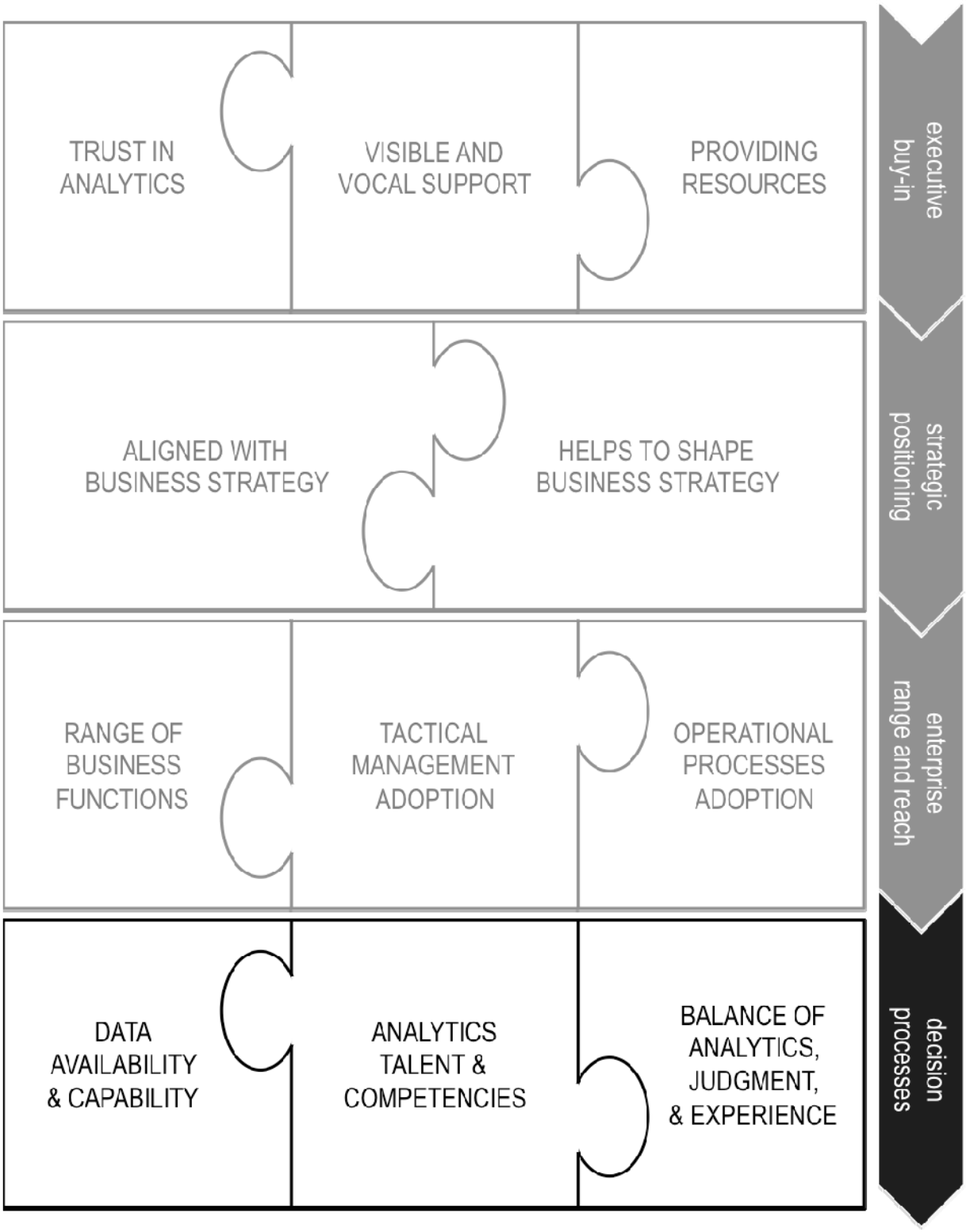
The facing page lists several practical guidelines to build useful models that have real business impact.

Module 4

Human Factors in Predictive Analytics

Analytic Culture

Decision Processes



Analytic Culture

Decision Processes

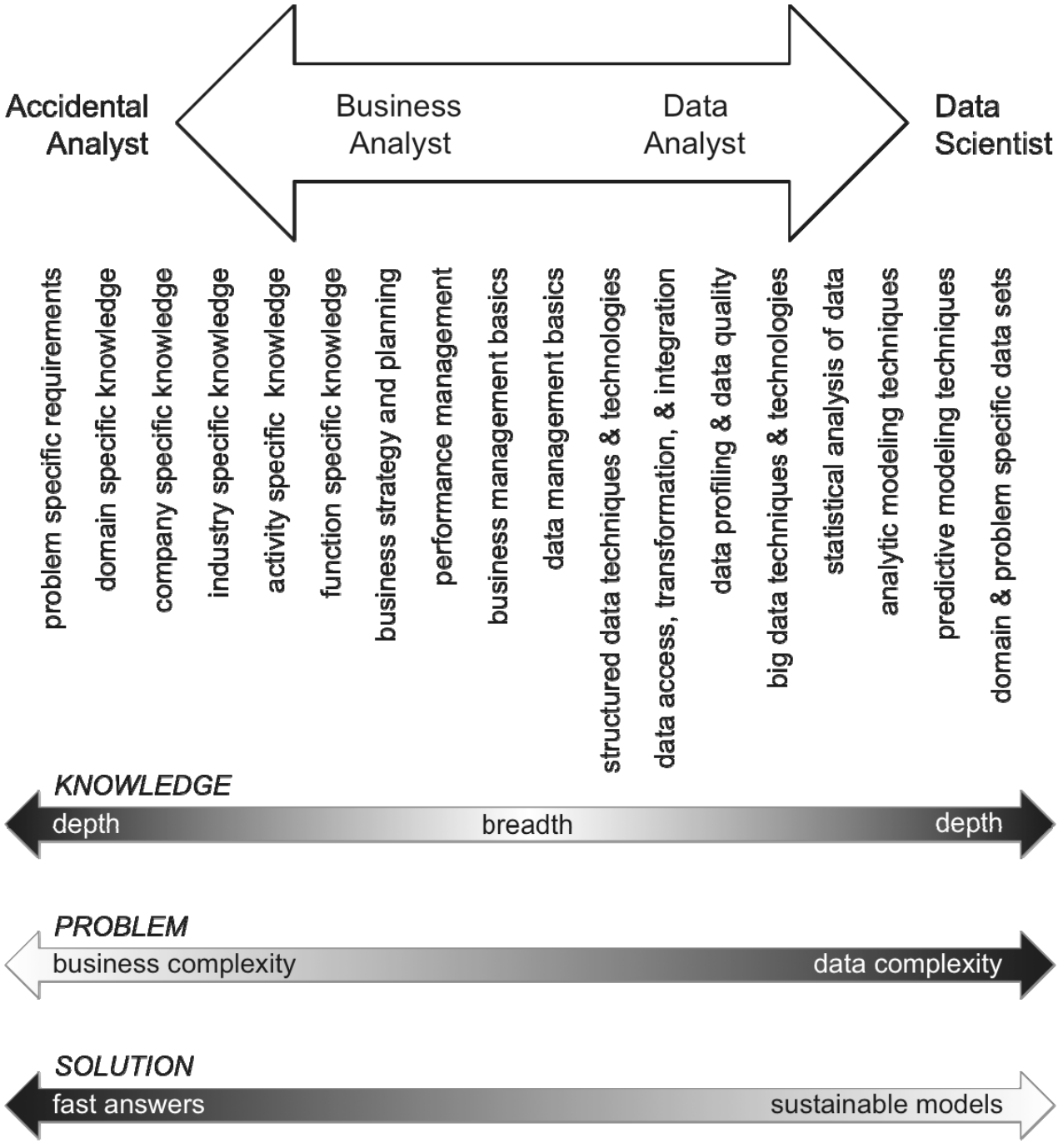
INFORMING DECISIONS

Ultimately, analytic culture has significant influence on decision making in business. A good goal for analytics-oriented organizations is to discount the popular phrase “data-driven decisions” and instead to strive for “analysis-informed decisions.” The fundamentals for a culture of analysis-informed decisions include

- available data
- data mangement capabilties
- people with analytics talent and competencies
- decision makers who blend analytics with experience and judgment for good decision making

People and Predictive Analytics

The Range of Knowledge



People and Predictive Analytics

The Range of Knowledge

APPLYING KNOWLEDGE TO ANALYTICS

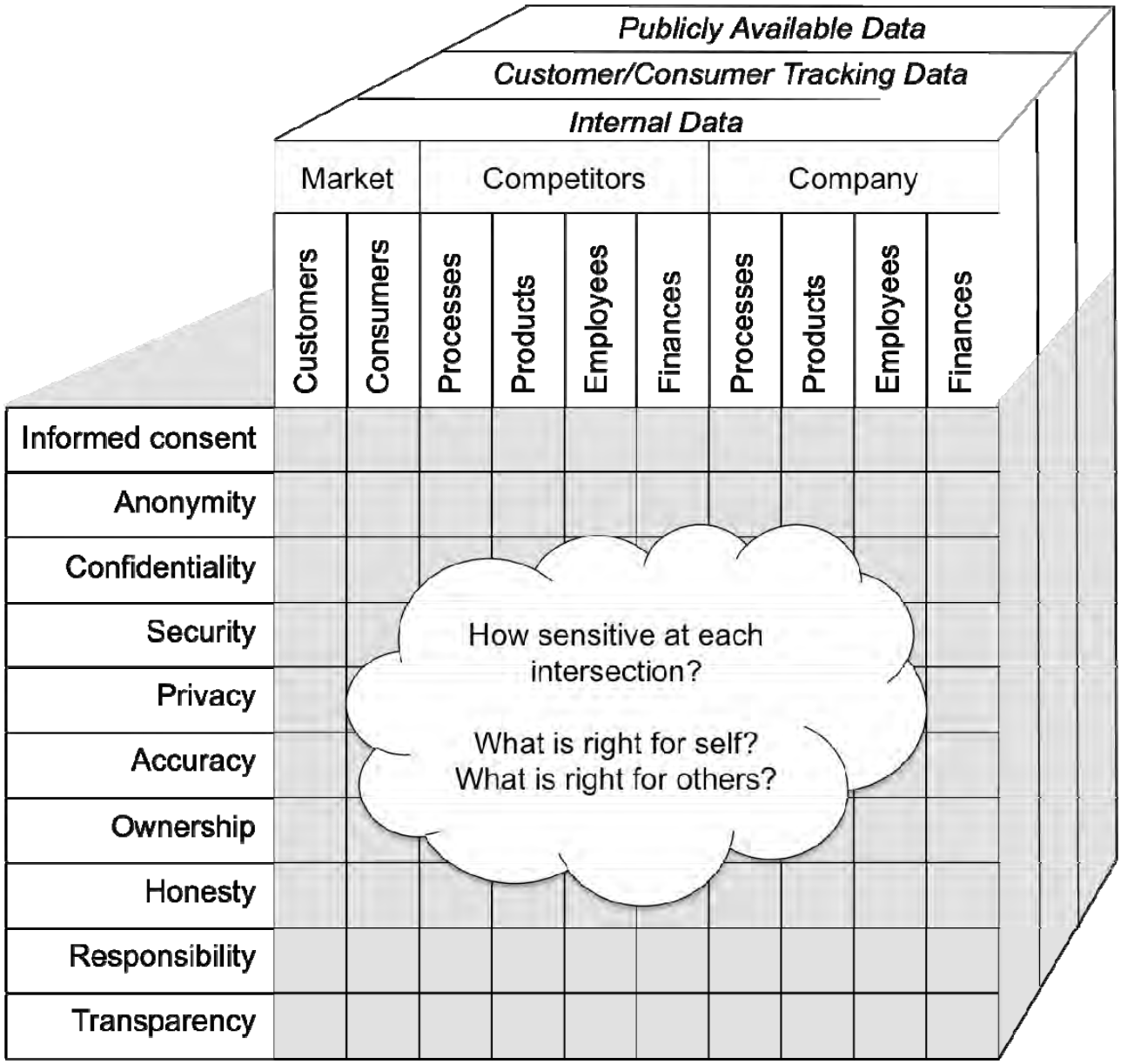
People are more than just titles, roles, responsibilities, and headcount. They have knowledge both individually and collectively, and the entire range of knowledge from depth of business knowledge to depth of technical knowledge is important to success with predictive analytics. One of the keys to success is ability to apply the right knowledge for the circumstances and specifics of a particular analytic problem.

DESKTOP VS. CENTRALIZED ANALYTICS

The discussion of where analytics belongs in an organization – with accidental analysts, with data scientists, or somewhere in the middle – doesn't have a single answer or a simple answer. A good guideline is to lean toward the data science end of the spectrum when data complexity is high and models need to be sustainable over a long lifespan. Lean toward accidental analysts when fast answers are more important than sustainable models and business complexity is greater than data complexity.

Ethics and Predictive Analytics

Data and Ethics



Ethics and Predictive Analytics

Data and Ethics

DATA AS INTELLIGENCE – COLLECTION, CREATION, AND PROTECTION

Ethics questions in predictive analytics center around the ways that we collect, manage, use and protect data, and they need to be considered beyond internal data – applied to publicly available data and especially for customer and consumer tracking data. The right-for-others vs. right-for-self considerations include customers and consumers, but also competitors and internal people, processes, and outcomes. Within this framework, key data-related questions include:

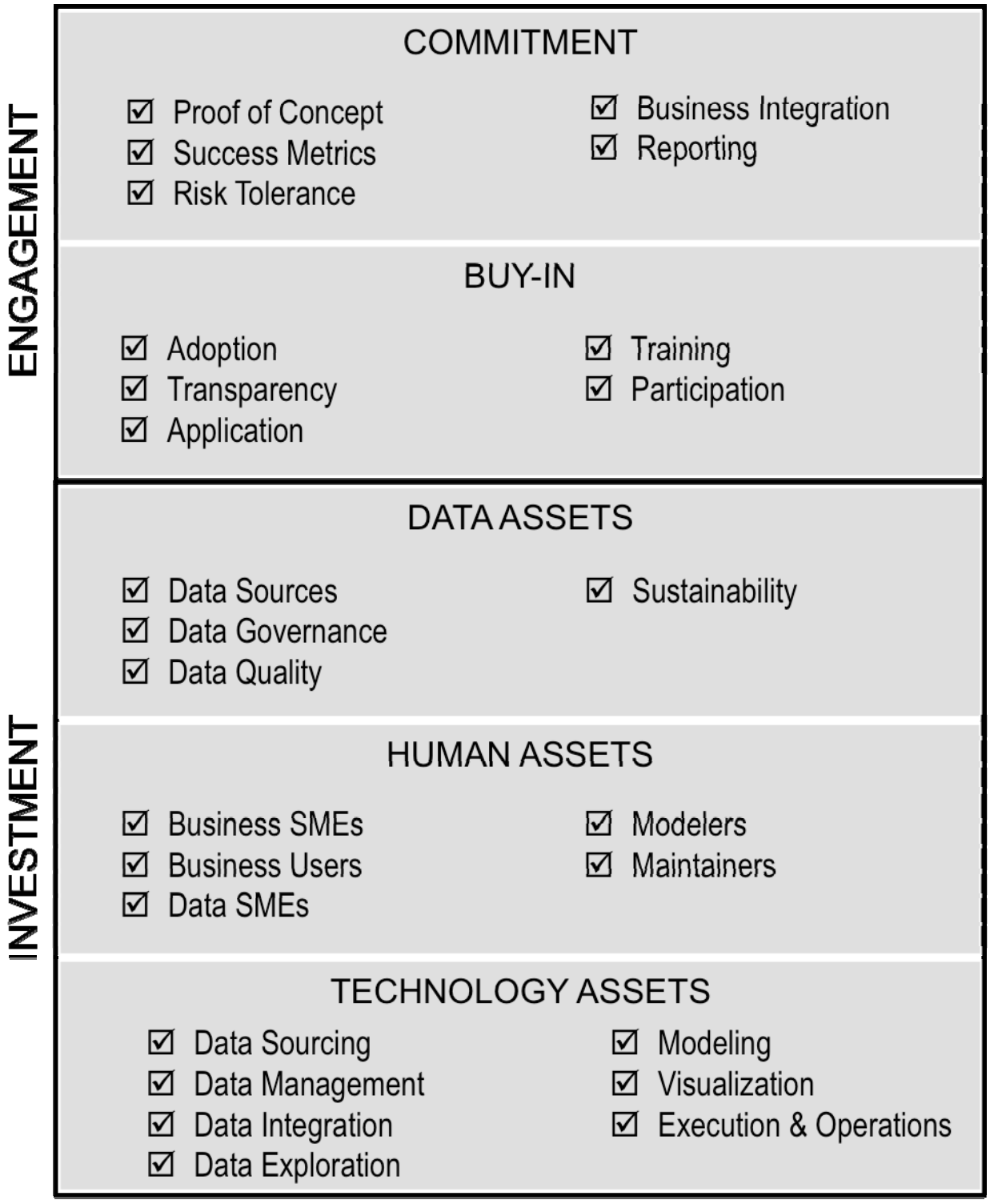
- Informed consent – Should the subject of the data know that data about them is collected, and should their agreement to the data collection activity be required?
- Anonymity – Should all personally identifying information be eliminated from the data? Should data be collected only in the form of aggregates such that individuals can't be identified?
- Confidentiality – Should sources and providers of data be protected from disclosure?
- Security – To what degree should data be protected from intrusion, corruption, and unauthorized access?
- Privacy – Should each individual have the ability to control access to personal data about themselves?
- Accuracy – What level of correctness is required of the data?
- Ownership – Is personal data about individuals an asset that belongs to the business or privately owned information for which the business has stewardship responsibilities?
- Honesty – To what degree should the business be forthright and visible about data collection practices?
- Responsibility – Who is accountable for use and misuse of data?
- Transparency – On a continuum with polar extremes of "totally open" and "stealth data collection" what is the right level of transparency

Module 5

Getting Started with Predictive Analytics

Predictive Analytics Readiness

Readiness Checklist



Predictive Analytics Readiness

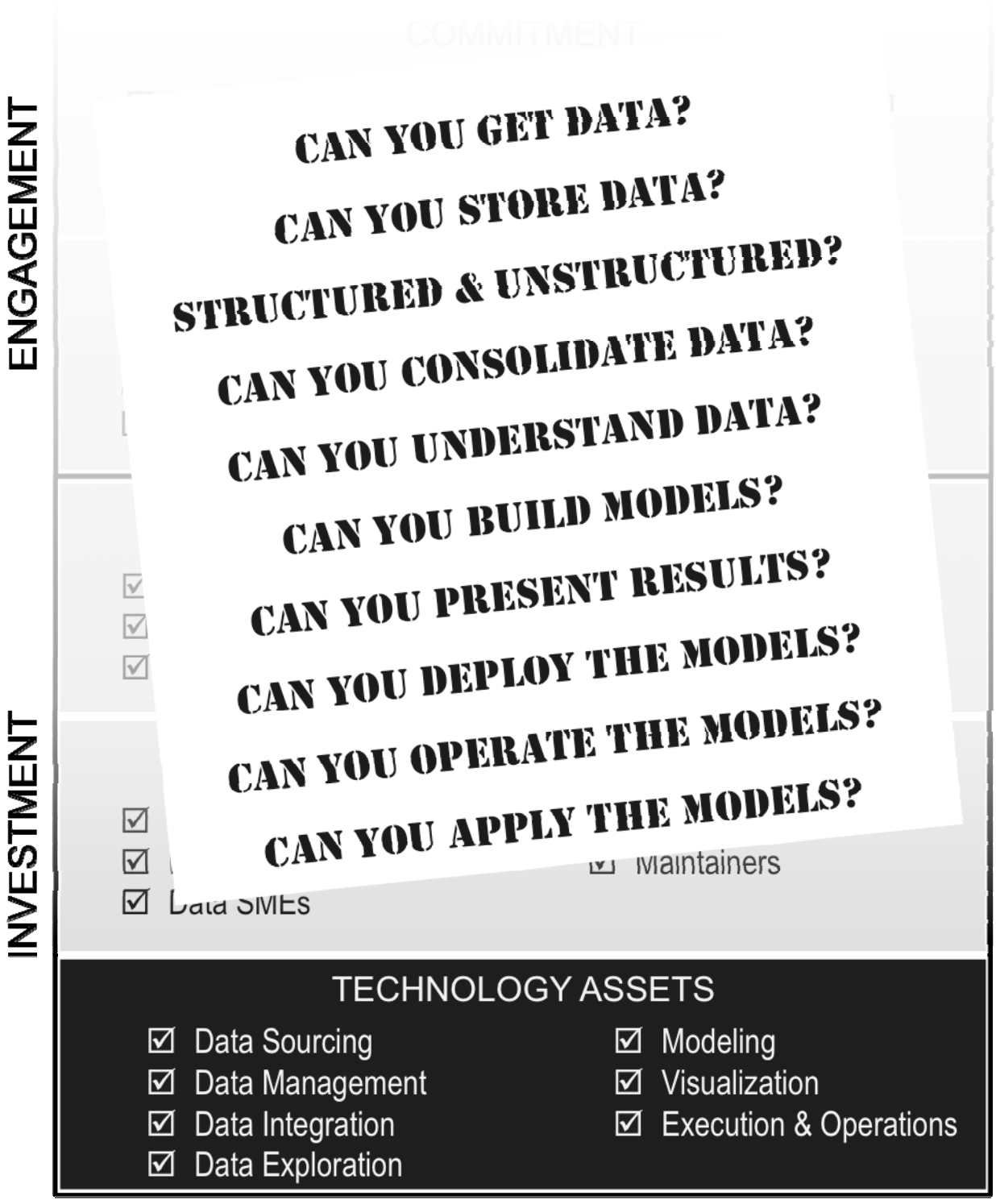
Readiness Checklist

ASSESSING THE CURRENT STATE

Think of predictive analytics as a journey along the path of maturing organizational intelligence and decison 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 Readiness

Technology Assets



Predictive Analytics Readiness

Tehnology Assets

TOOLS AND TECHNOLOGY

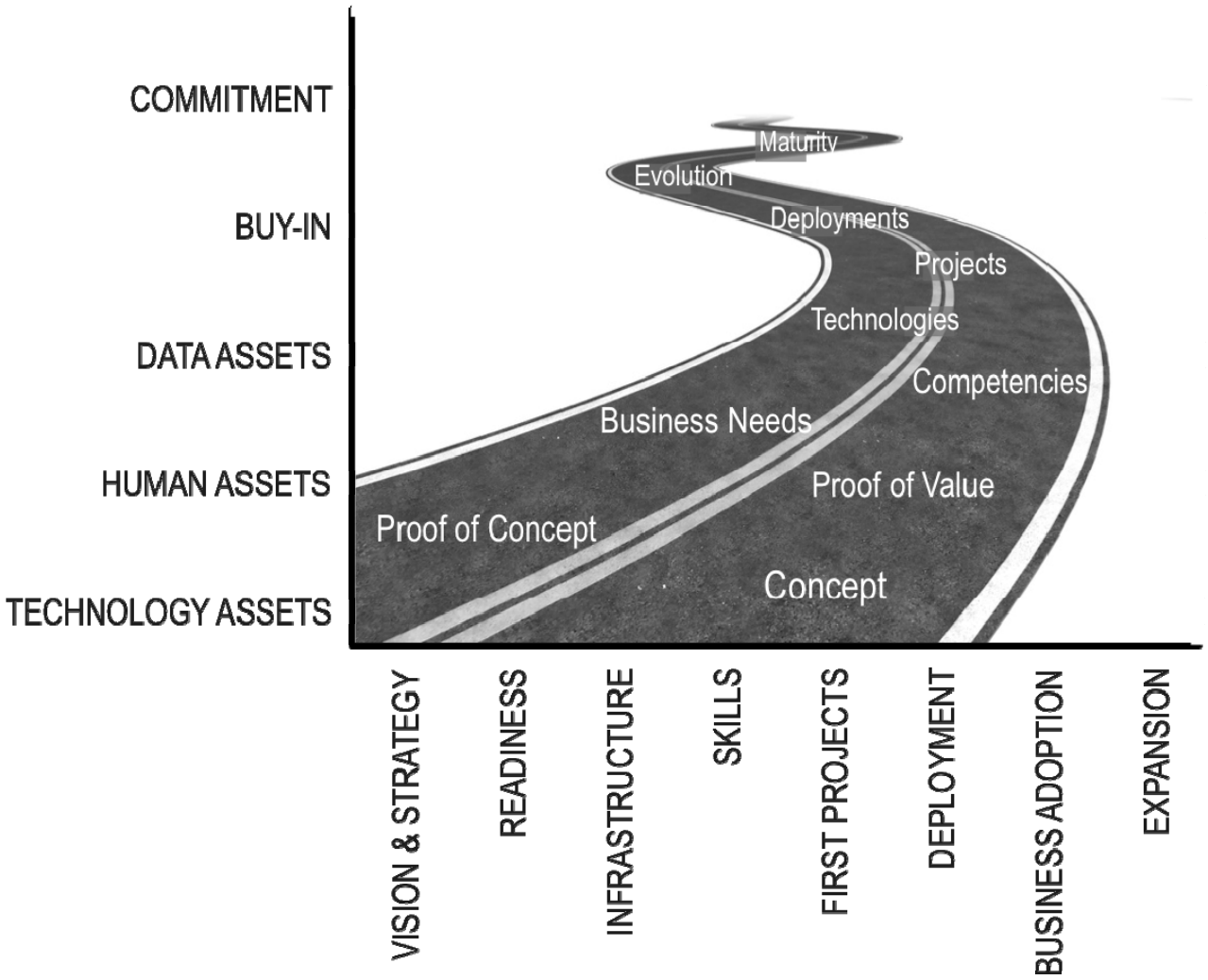
Technology is fundamental to predictive analytics. Making sure that you have the right technology assets is part of any readiness assessment. Consider each of these technology categories:

- Data Sourcing – the technology needed to get data
- Data Management – technologies for storing, formatting, and preparing data
- Data Integration – technologies to consolidate data from multiple sources
- Data Exploration – technologies to profile and understand data
- Modleing – technologies to build mining and analytic models
- Visualization – technologies for charting, graphing, and data presentation

For each technology category evaluate what you have, what you need, what gaps exist, and how you will fill those gaps.

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, activities, and to help meet those goals in an organized and step-by-step way. The roadmap has three purposes:

- Buidling 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 learing from first projects
- Deploying business solutions
- Growing business adoption
- Expanding solutions, adoption, and analytic maturity

