Transforming Data
With Intelligence™

TDWI Analytics Principles and Practices

Delivering Business Insight from Data

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

tdwi

Transforming Data
With Intelligence™

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. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY.

COURSE OBJECTIVES

You will learn:

- The concepts and practices of analytic modeling
- Fundamentals of data literacy
- An analytics topology to make sense of the variety of analytics types and techniques
- The data side of analytics including data sourcing, data discovery, data cleansing, and data preparation
- Analytics techniques for exploration, experimentation, and discovery
- The human side of analytics: communication, conversation, and collaboration
- The organizational side of analytics: self-service, central services, governance, etc.
- A bit about emerging techniques and technologies shaping the future of analytics

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 2020

© Copyright 2020 by TDWI, a division of 1105 Media. All rights reserved. No part of this document may be reproduced in any form, or by any means, without written permission from The Data Warehousing Institute.

TABLE OF CONTENTS

tdwi

Transforming Data With Intelligence™

Module 1

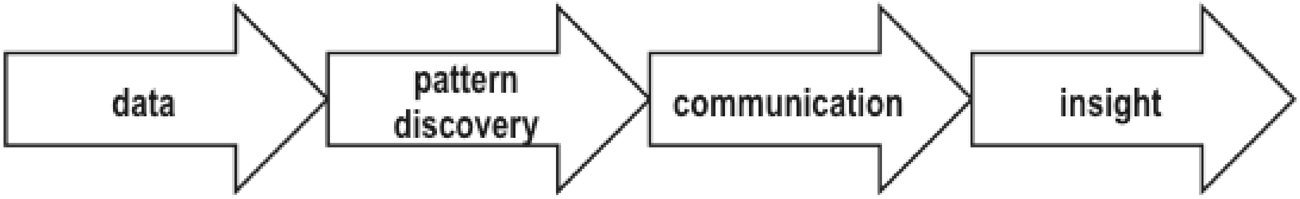
Analytics and Data Literacy Concepts

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Defined

From Data to Insight

> Analytics is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight.
>
> wikipedia.org



statistics … quantification … programming … visualization

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Defined

From Data to Insight

WHAT IS ANALYTICS?

The facing page shows Wikipedia's definition of analytics with focus on data, patterns, visualization, and insight. The goal of analytics is to find insight by examining data. Key concepts in getting from data to insight include pattern discovery, quantification, statistical analysis, and visual communication.

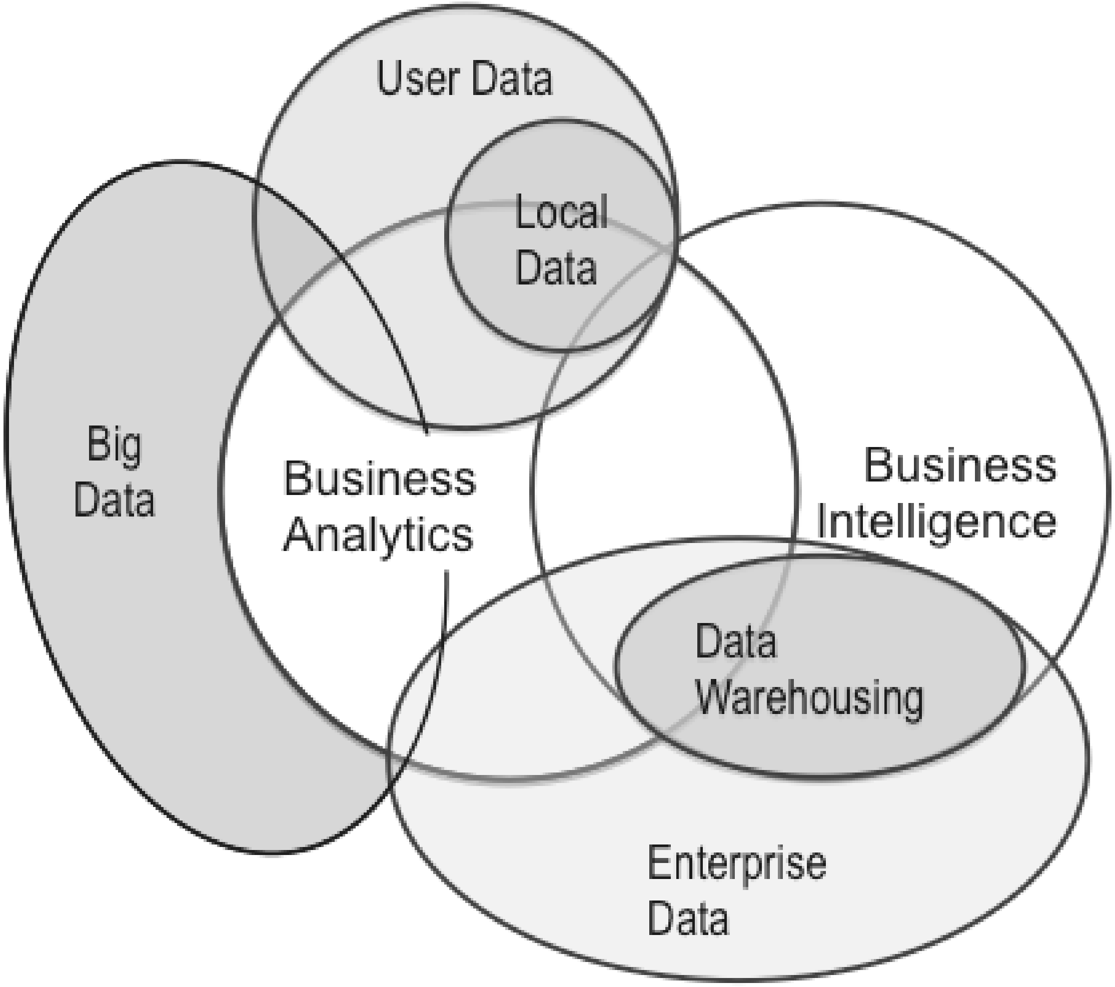
IT TAKES PEOPLE TOO

The definition concentrates on the computer enabled aspects of analytics. But it is important to remember that analytics involves more than data and computer processing. Analytics uses "data visualization to communicate insight" *to people.* People combine the insights of analytics with their knowledge, experience, and judgment to interpret meaning and to make decisions.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Data Literacy Foundations

Data



Scope of Data
Finding Data
Observations & Populations
Raw Data vs. Summary Data
Data Preparation

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Data Literacy Foundations

Data

DATA LITERACY

Data literacy skills are crucial tools for driving business impact through analytics. Whether you are a business practitioner, an analyst, or a data scientist, you must understand how to think about data resources and apply them to solve business challenges. Foundational data literacy concepts include the description and classification of data, methods for manipulating data to generate insight, and communication skills for persuading business people to take action.

SCOPE OF DATA

Analytics makes extensive use of data, both quantitative and qualitative. Quantitative data uses numbers to express business events, behaviors, and trends as measures. Qualitative data segments data instances by categories. Qualitative data is often referred to as categorical data. Both quantitative and categorical data have roles in statistical analysis.

Analytics data comes from many sources. Unlike BI, which primarily focuses on enterprise data and the data warehouse, the scope of analytics data is quite broad including:

- User data – the data found in departmental and end-user databases – is valuable and commonly used in analytics.
- Local data is a distinct subset of user data that is often found in spreadsheets. It may be maintained locally to meet individual needs, downloaded from a warehouse and then manipulated, created manually to meet a specific need, acquired or derived from external data sources, or generated by earlier analytics processes.
- Big data offers a variety of data sources to enrich the analytics process and expand analysis opportunities including data from web searches, online shopping, email, text messaging, social media activity, machine-to-machine communications, sensor data, and much more.
- Enterprise data that is widely used across multiple business functions and is defined, managed, and governed from a global or enterprise-wide perspective.
- Warehouse data this integrated data from multiple enterprise sources, removing redundancy and anomalies and standardizing data representation.

Transforming Data With Intelligence™

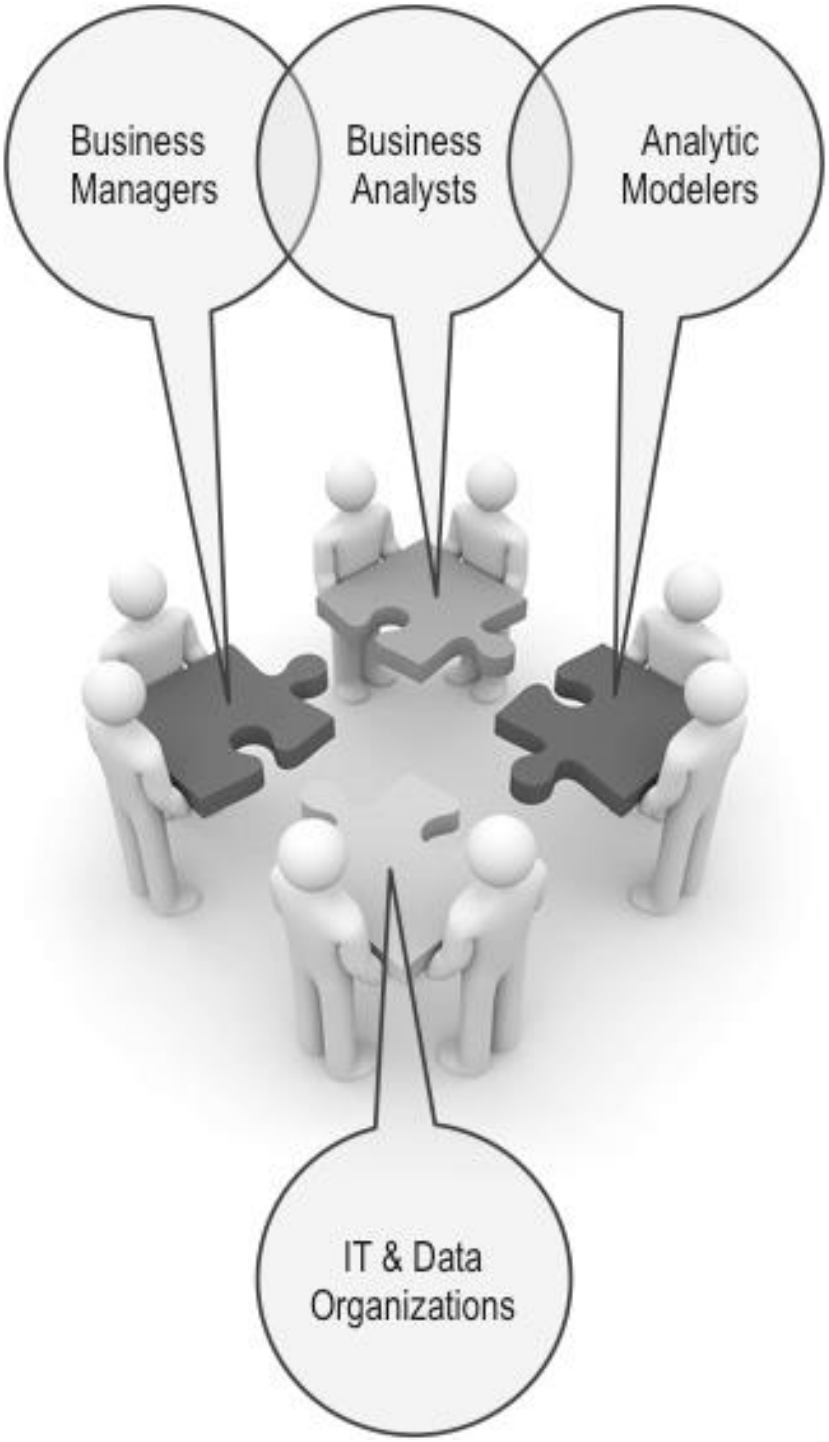
Module 2

The Analytics Environment

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Stakeholders

The Participants



© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Stakeholders

The Participants

ANALYTICS STAKEHOLDERS

Four groups of people have roles, responsibilities, and a stake in business analytics:

- Business managers have a primary role in analytics. They are the decision makers and the planners who need to gain insight and reduce uncertainty through analytics. These people have critical responsibilities in framing and defining problems.

- Business analysts are the people who analyze business behaviors. They are responsible for evaluating and interpreting the results of analytic models – to develop conclusions, identify alternatives, and make recommendations. In many organizations there is a high level of overlap between business managers and business analysts. The managers often perform their own analyses.

- Analytic modelers are the people who analyze data to identify business behaviors. They are responsible for understanding analytics needs or problems, obtaining and preparing data, and applying statistical methods to derive information and insight from the data. Analysts and modelers often overlap.

- Data and IT organizations (e.g., data stewards, data governance council, etc.) have the responsibility to provide some of the data that is needed for analytics, to know where and how enterprise data is used, and to secure privacy-sensitive data from unauthorized access and use.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Organizations

Organization Models

SELF SERVICE for autonomy and ability to quickly meet needs of individual business units

SHARED SERVICES to propagate standards and best practices across business units and for time, cost, and resource efficiencies

CENTRAL SERVICES for high levels of control, governance, standardization, and consistency across all business units

HYBRID SERVICES to adapt to the diverse kinds of analytics requirements and projects that are sure to occur throughout the organization

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Analytics Organizations

Organization Models

SELF SERVICE

The self-service model creates an environment where business units meet their own analytics needs with support of business-oriented tools, architectures, frameworks, guidelines, examples, templates, etc. This model is suited to well-defined problem domains where business users have a desire for autonomy and a relatively high level of data analysis skills.

SHARED SERVICES

The shared services model defines processes, standardizes architecture, and maintains a centralized team for shared work, but most project and process work occurs in individual project teams and distributed lines of business. The blend of centralized and decentralized resources achieves good efficiency of resource utilization. The centralized team is focused primarily on critical skills and on those shared resources where no single project has fulltime needs.

CENTRAL SERVICES

In the central services model, standards, processes, architecture, and technology are prescribed. A single, centralized team is responsible for development, deployment, and management of analytics solutions. This model works well when goals are exceptional consistency, strong governance, rapid delivery, and managed costs. In an environment of high demand, the central services model may be challenged to scale up to meet demand.

HYBRID MODELS

As a practical matter, many organizations evolve to a mix-and-match hybrid of the three service models. Good guidelines and clear understanding of the criteria by which projects and service models are matched is important to avoid misuse of any of the service levels.

Transforming Data
With Intelligence™

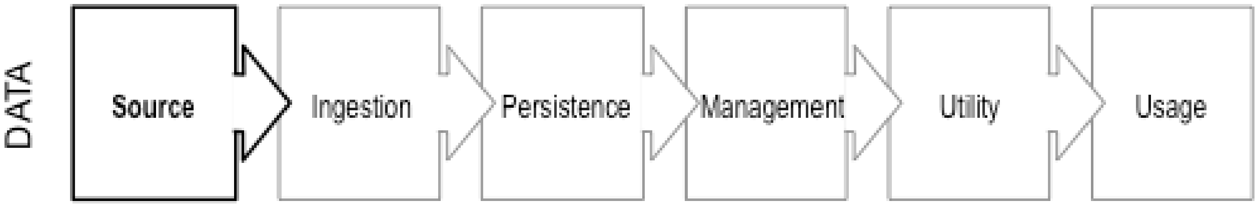
Module 3

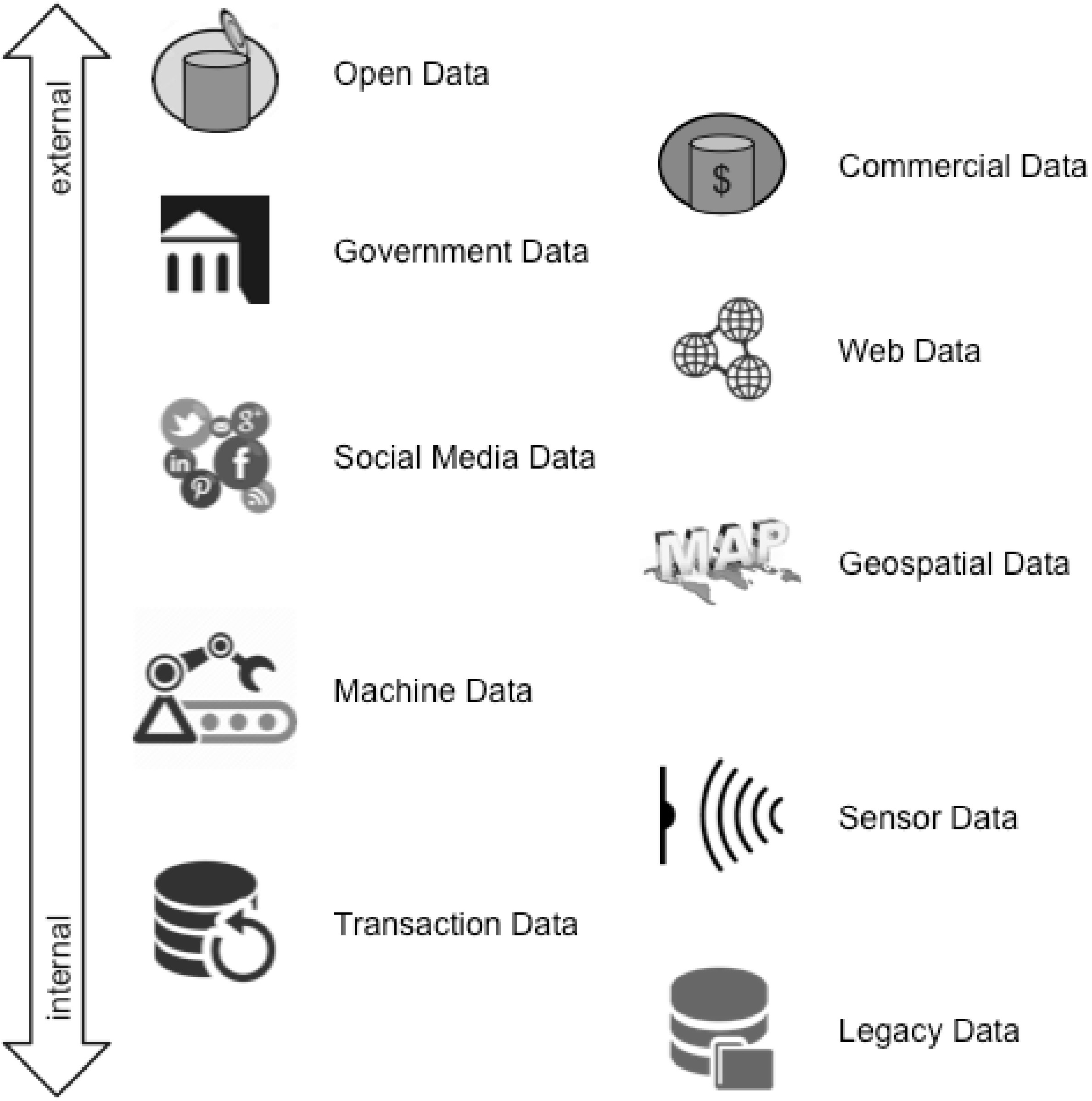
Analytics Architecture

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Data Architecture

Data Sources and Types





Data Architecture

Data Sources and Types

INFORMATION RESOURCES

The data architecture defines standards and guidelines for the acquisition, management, and consumption of information resources in the analytics program.

CATEGORIES OF DATA SOURCES

The available data for analysis is diverse and rich with opportunities. As an architectural component, source data describes the available data in terms of origin, context, structure, and latency. Data sources can be classified according to the following dimensions.

Data Origination

- Internally located within the enterprise
- Externally located outside of the enterprise

Data Context

- Open – licensed for free reuse
- Commercial – purchased under contract
- Government – available from government agencies
- Web – general availability by web scraping
- Social Media – using interfaces from providers
- Geospatial – relates to mapping and locations
- Machine – generated by devices, equipment & machinery
- Sensor – measurement oriented and continuous
- Transaction – event oriented and discrete
- Legacy – available data from older technology (online or offline)

Data Structure

- Structured data – fields are organized into a tabular format
- Unstructured data – cannot be stored as fields in a tabular format

Data Latency

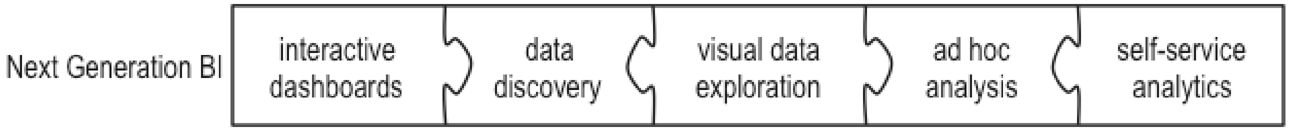
- Real Time – data latency is close to zero
- Near Time – data latency is less than 24 hours
- Off Line – data latency is greater than 24 hours.

Describing data in terms of these four dimensions and the categories with each dimension helps to understand and manage the diversity of data that is available in today’s analytics world.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Process Architecture

Next Generation BI



© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Process Architecture

Next Generation BI

WORKING WITH INFORMATION

Process architecture describes key functionality needed to enable analytics and business capabilities delivered by business analytics.

ADVANCES IN BUSINESS INTELLIGENCE

BI has evolved significantly since it was originally defined in the 1990s. It was initially defined as a passive system predominately delivering static information using reporting and OLAP tools.

BI is now recognized to be an active system that delivers functionality in addition to information. A variety of new innovations enable these advances by transforming BI from a passive to a functional system.

The following items are examples of advances in BI that enable components of the process architecture.

- Interactive Dashboards – evolution from static reports and displays to delivering interactive user experiences in a metrics-oriented environment

- Data Discovery – searching and exploring data sets to identify patterns and relationships either visually or with machine-learning assistance.

- Visual Data Exploration – use of interactive graphical representations of data values to visually identify patterns and relationships.

- Ad Hoc Analysis – process of business analysts framing an assigned problem and interactively determining and acquiring the data needed to analyze and solve the problem.

- Self-Service Analytics – business analysts or managers interactively determine what measures and metrics are needed to answer what, why, and what-if questions important to management decision making. Business people, without IT intervention, acquire data, prepare data, and visualize and analyze interactively using business-friendly tools that require little or no coding or technology expertise.

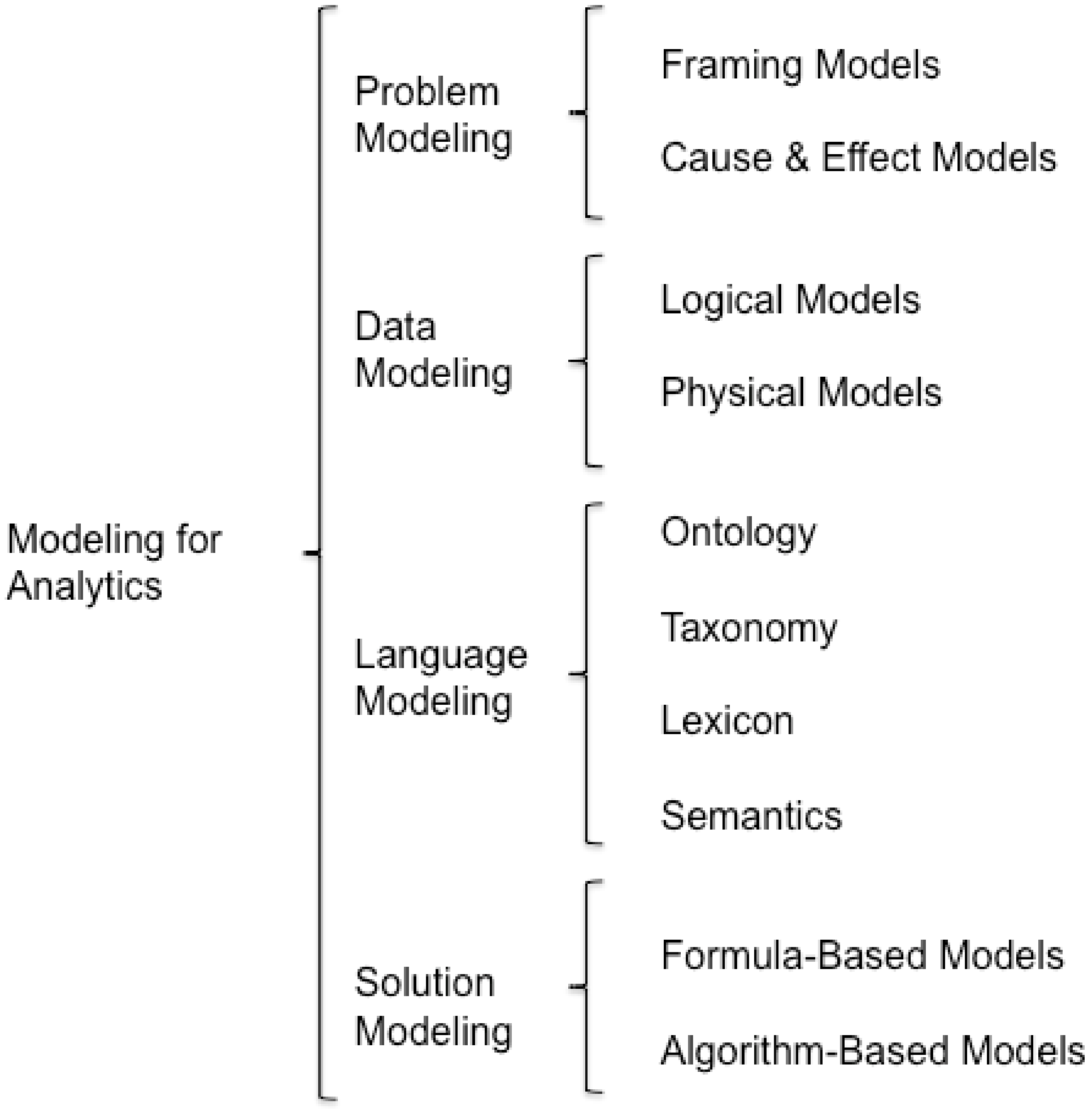
Module 4

Analytic Modeling

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

The Roles of Models

Why Model?



© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

The Roles of Models

Why Model?

WHAT IS A MODEL?

A model is an abstract representation of something in the real world. Models help us to understand complex things by viewing them at varying levels of abstraction and from multiple perspectives. Most frequently, when people use the term "analytic modeling" they refer to the process of building solution models. The full scope of modeling for analytics is much broader, including problem models, data models, language models, and solution models.

PROBLEM MODELING

Any analytics effort begins by understanding the problem space. This course examines two kinds of problem modeling – one for problem framing and one for cause-effect modeling.

DATA MODELING

Analytics is a data-driven activity, so understanding of the data is essential to analytics processes. Data models are an effective way to examine and document the content, structure, and relationships that exist in a set of data. Both logical models that describe data in business context and physical models that describe technical implementation may be useful. Models may be developed to prescribe schema for data storage (schema on write for relational tables) or to describe implied schema for data consumption (schema on read for NoSQL). We'll look briefly at data models, but this course does not explore them in depth.

LANGUAGE MODELING

When data for analytics includes text, analysis processes must parse and inspect text to turn it into useful and quantitative data. Text analysis uses four kinds of models: Ontological models are a linguistic structure to describe things in the real world and how they are related. Taxonomic models describe hierarchical relationships (parent/child) in a classification structure. Lexical models describe the meaning of terms in a specific domain – a single term, for example, may have an entirely different meaning in financial services than in healthcare. Semantic models describe the organization of words into sentences. They are used to accurately parse sentences and find the meaning in them. We'll look briefly at language models, but this course does not explore them in depth.

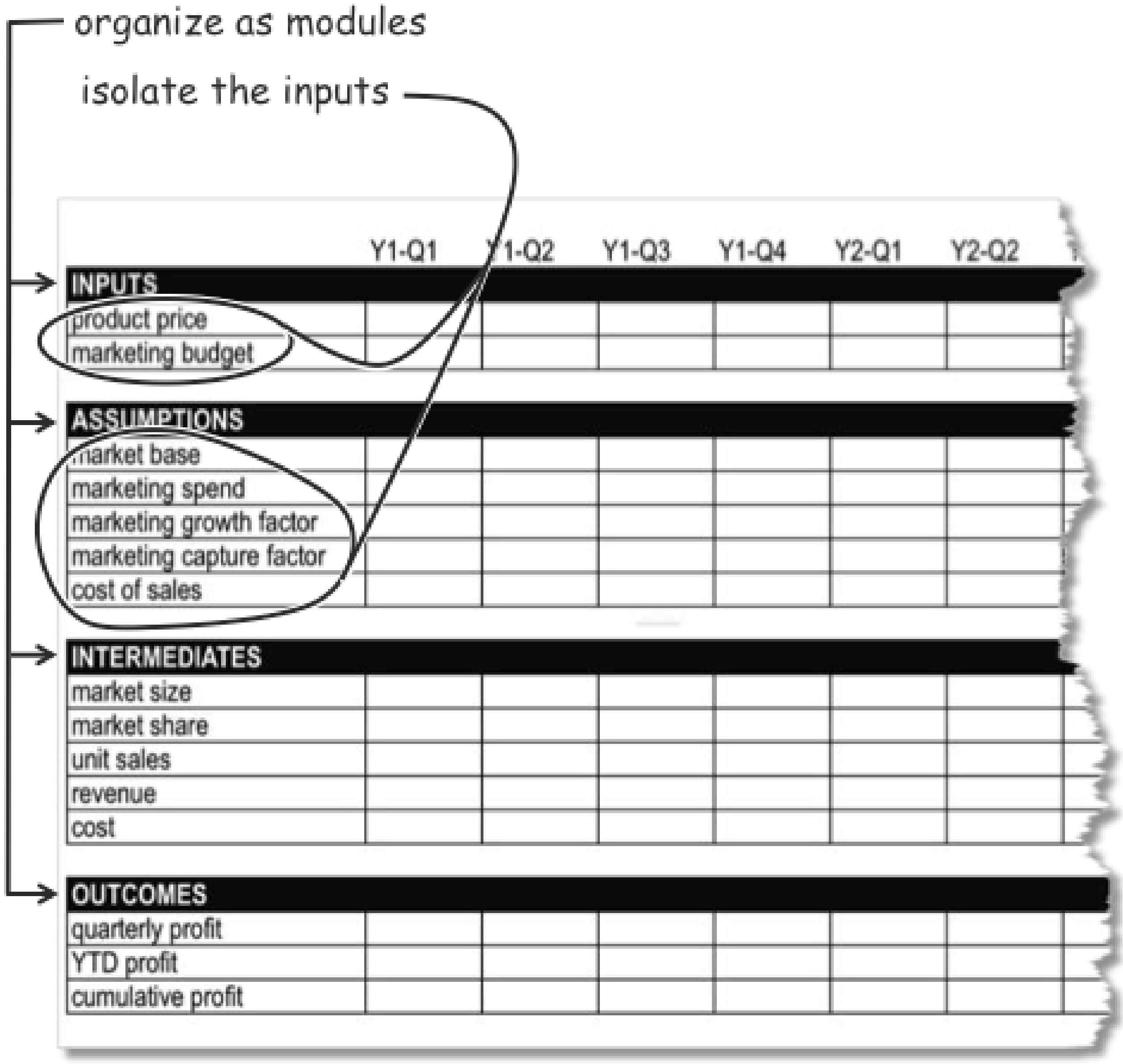
SOLUTION MODELING

Solution models are based on understanding of business dynamics – when *X* occurs, *Y* reacts in this way. The most common solution models are of two types: formula-based and algorithm-based. We'll look at both types and have opportunity to practice formula-based modeling.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Solution Modeling

Formula Based Modeling – Structuring



	Y1-Q1	Y1-Q2	Y1-Q3	Y1-Q4	Y2-Q1	Y2-Q2
INPUTS						
product price						
marketing budget						
ASSUMPTIONS						
market base						
marketing spend						
marketing growth factor						
marketing capture factor						
cost of sales						
INTERMEDIATES						
market size						
market share						
unit sales						
revenue						
cost						
OUTCOMES						
quarterly profit						
YTD profit						
cumulative profit						

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Solution Modeling

Formula Based Modeling – Structuring

SPREADSHEET ENGINEERING

The technique of "spreadsheet engineering" is used in this course to illustrate many techniques of modeling analytics solutions. It is not recommended, nor is it practical, to meet all of your analytics needs with spreadsheets. Yet there are many good reasons to take a spreadsheet view:

- Much of business analysis, especially the analysis performed by business managers, is done with spreadsheets.
- Regardless of the analytics tool that you use, you will work with data that is organized in rows and columns and that has relationships among the cells.
- The kinds of variables illustrated with spreadsheets – inputs, assumptions, intermediates (unknowns), and outcomes – apply for every solution modeling problem and every analysis tool.

WHY ENGINEERING?

It may sound like an ominous term – spreadsheet engineering – but the real goal is to plan and design before building. All too often the initial form of a spreadsheet is determined by the source data that is available. We load the data and that determines the rows and columns. Then we take a circuitous path of fit-and-fix, poke-and-patch until we arrive at something close to a desired solution. A better alternative is to begin at the end – to start with the desired outcome and follow the chain backward to the inputs, carefully managing data relationships and dependencies along the way.

THE BASICS

Begin with a quick sketch of the spreadsheet that separates components into modules that are "logically, physically, and visually distinct."[1] The modules may be sections within a worksheet as shown here, or they may be separate worksheets in a workbook for more complex models. The elements of an influence diagram – decision variables, deterministic variables, chance variables, and outcome measures – are a good first cut at modularity.

Use modularity to isolate the input variables. All of the numerical inputs to the model – decision variables, deterministic variables, and assumptions – should be grouped together and modularized. Ideally the inputs are placed at the top of the worksheet and dependencies cascade downward.

[1] *Modeling for Insight*, pp. 37-38, Powell and Batt

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

tdwi

Transforming Data
With Intelligence™

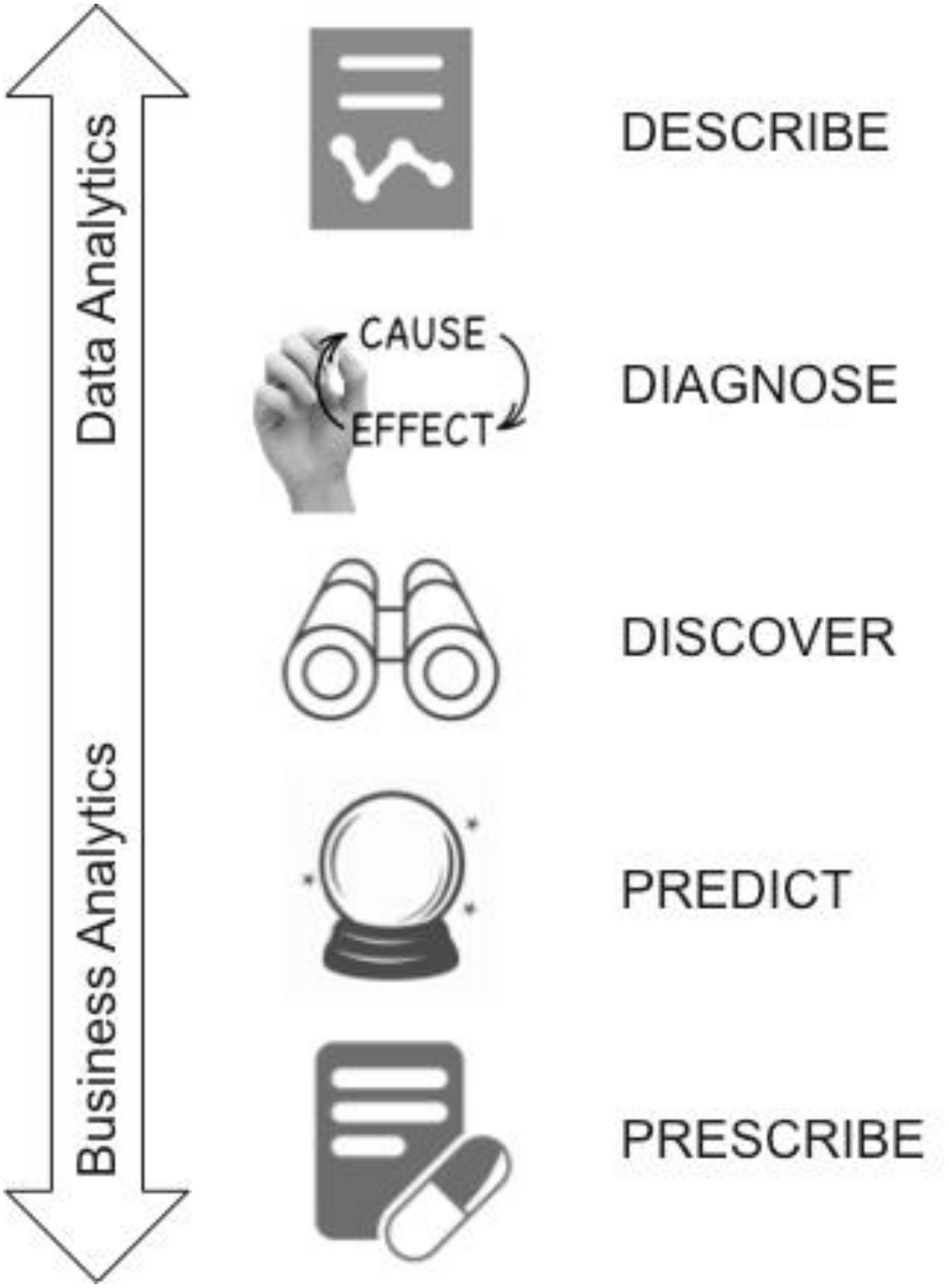
Module 5

Applied Analytics

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Five Kinds of Analytics

What We Do



© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Five Kinds of Analytics

What We Do

DIFFERENCES IN ANALYTICS FOCUS

Analytics is not a “one size fits all” endeavor. Five kinds of analytics are commonly practiced – each with a different purpose.

- Descriptive and diagnostic analytics are largely data focused and seek understanding of past events.
- Discovery analytics builds the bridge from data focused to business focused, looking at data-to-business connections.
- Predictive and prescriptive analytics, though data dependent, are very much business oriented and looking to the future.

ANALYTICS DEPENDENCIES

Though not a hard-and-fast rule, there are some dependencies among the types of analytics. Diagnostic work is difficult without first performing descriptive work to understand the nature of the data and the events that it describes. Discovery analytics benefits from results of descriptive and diagnostic findings. Predictive modeling benefits from discovery, diagnosis, and description, and prescriptive analytics is built on a foundation of predictive analytics.

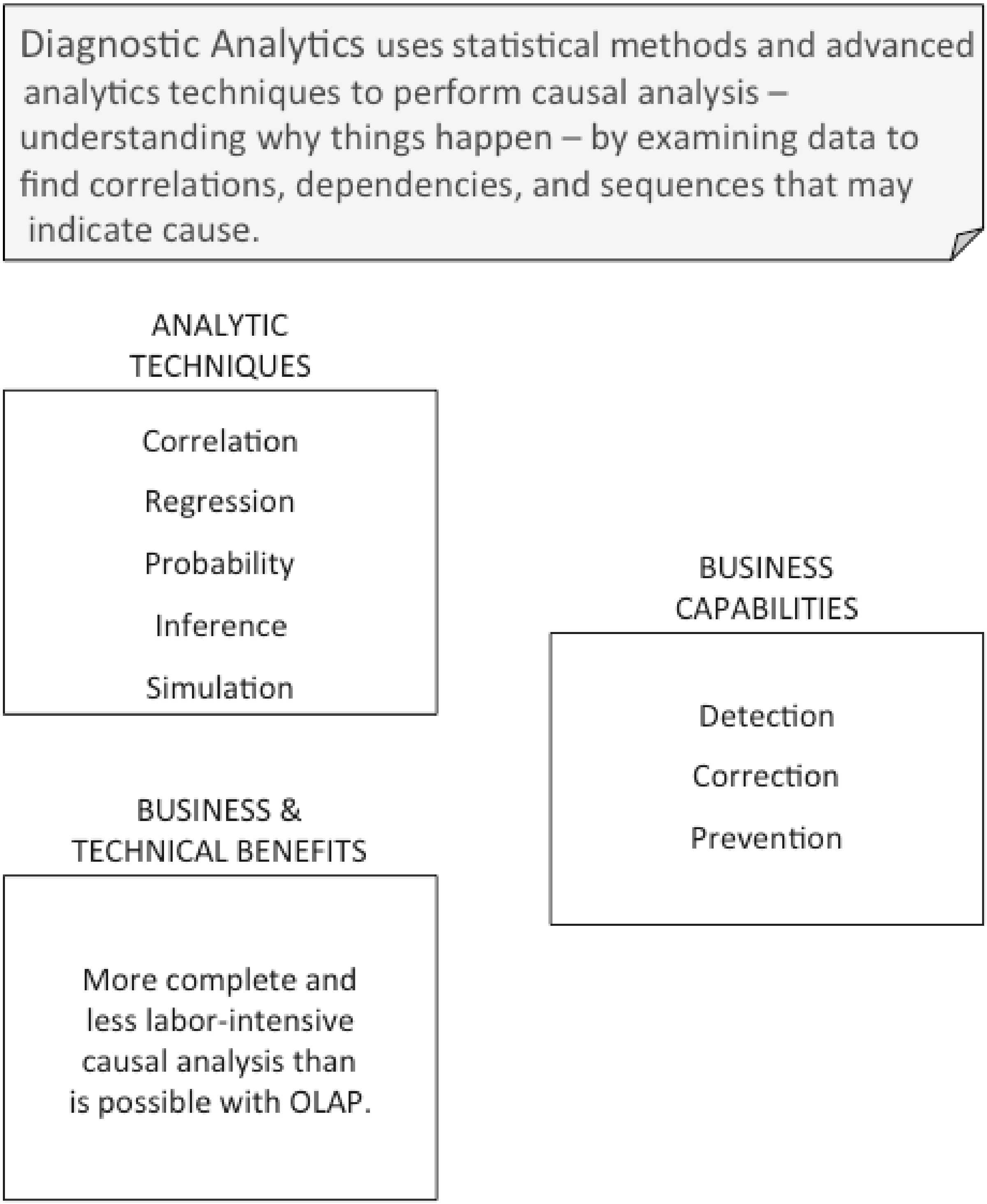
A WORD OF CAUTION

Consider the above description of dependencies to be a general guideline, and don’t be led to believe that analytics progression is a linear path. Analytics is always iterative, and dependencies can and do go both directions – up and down the chain. Discovery analytics, in particular, has bi-directional dependencies as it is applied both for data discovery and for business discovery.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Diagnostic Analytics

Definition and Description



Diagnostic Analytics uses statistical methods and advanced analytics techniques to perform causal analysis – understanding why things happen – by examining data to find correlations, dependencies, and sequences that may indicate cause.

ANALYTIC TECHNIQUES

- Correlation
- Regression
- Probability
- Inference
- Simulation

BUSINESS CAPABILITIES

- Detection
- Correction
- Prevention

BUSINESS & TECHNICAL BENEFITS

More complete and less labor-intensive causal analysis than is possible with OLAP.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY

Diagnostic Analytics

Definition and Description

WHY DID IT HAPPEN?

The purpose of diagnostic analytics is to detect unusual situations or abnormal conditions in the context of a business or operational process. This detection capability is combined with root cause analysis methods to gain insights into causal relationships driving the observed or predicted fault. It provides fault detection and root cause analysis capabilities to the different categories of people who need to respond to the situation and take corrective action.

Considering that the various types of analytics enable and support each other, this form of analytics depends on discovery analytics for base models and descriptive analytics for statistical properties of key processes. Diagnostic Analytics detects and identifies current unusual conditions. However, this capability can be combined with predictive and prescriptive analytics to create predictive diagnostics, identifying potential problems before they actually occur. Prescriptive analytics may apply to recommended action for either detected or predicted problems.

TECHNIQUES

Detection, correction, and prevention are the core sets of techniques that apply to diagnostic analytics. Specific detection techniques include:

- Statistical process control charting based on Shewhart and CUSUM (cumulative sum) charts
- Auto-regressive Models
- Logistic Regression Models
- Monte Carlo Simulation Models
- Probability Models
- Naïve Bayes Classification Models
- Linear Regression Models
- Process Simulation Models

Correction and prevention techniques are based in root cause analysis methods and include Five Why’s, fishbone diagramming, and causal loop modeling.

BUSINESS VALUE

The facing page itemizes the primary business capabilities that are enabled with diagnostic analytics, as well as key business and technical benefits. Note that diagnostic analytics provides efficiency above and beyond the OLAP slice-and-dice approach to causal analysis.

© TDWI. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. DO NOT COPY