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.

This page intentionally left blank.

TDWI Analytics Fundamentals

COURSE OBJECTIVES

You will learn:

- The concepts and practices of analytic modeling
- An analytics topology to make sense of the variety of analytic types and techniques
- The data side of analytics including data sourcing, data discovery, data cleansing, and data preparation
- Analytic 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 2016

TABLE OF CONTENTS

Transforming Data With Intelligence™

Module 1

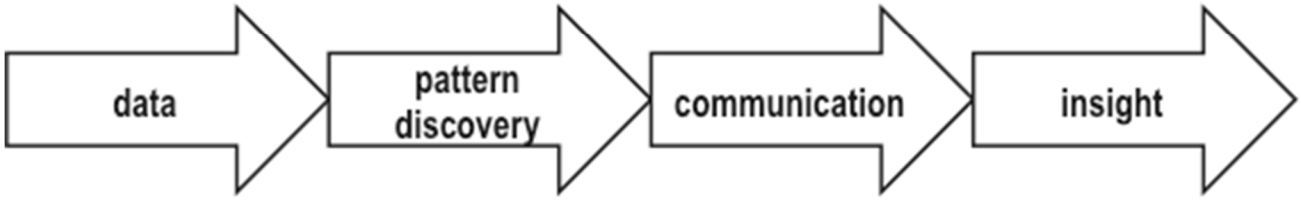
Concepts of Analytics

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

Analytics Defined

From Data to Insight

WHAT IS ANALYTICS?

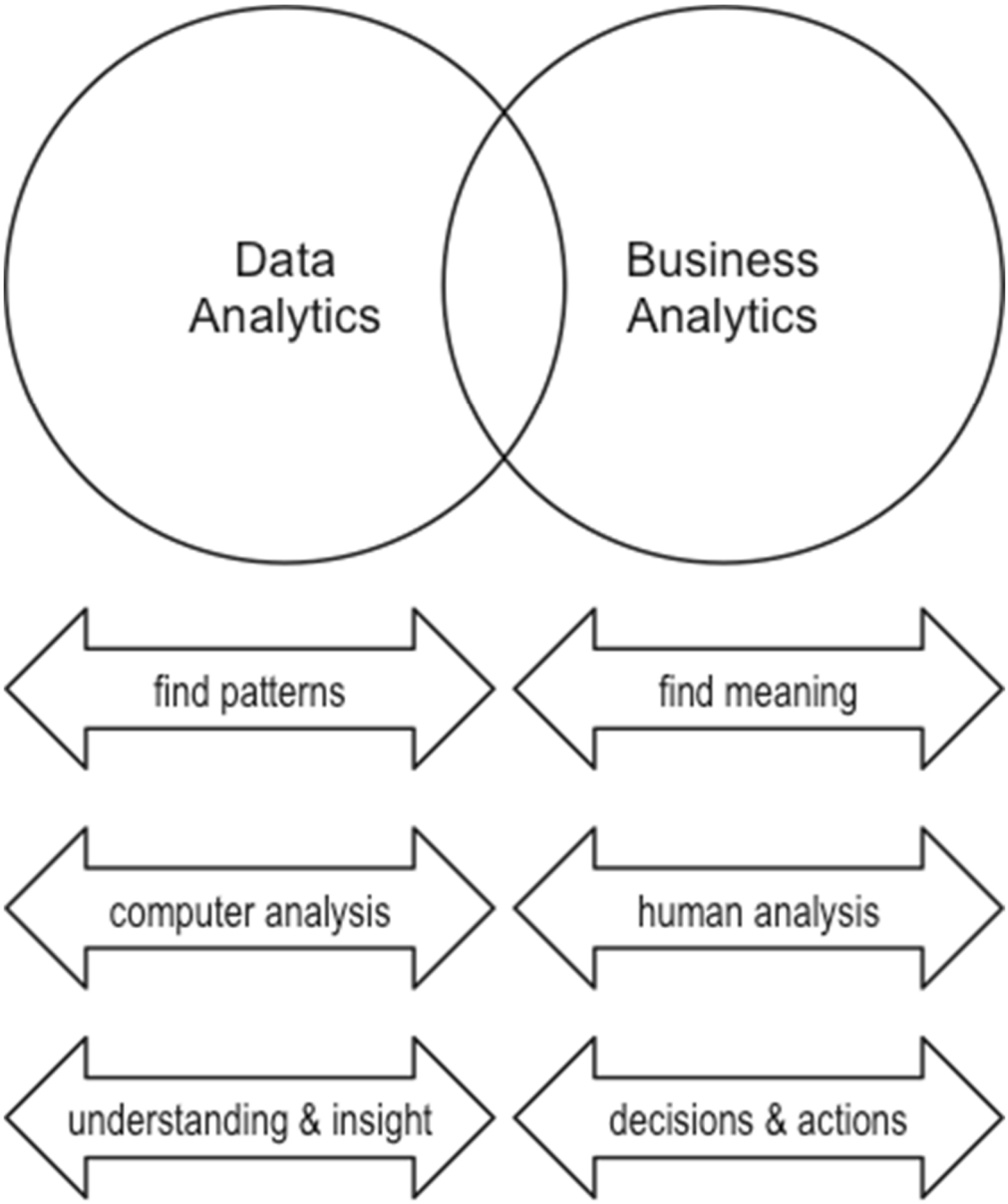
The facing page shows Wikipedia's definition of analytics with focus on data, patterns, visualization, and insight. The goal of analytics is to find insight by examining data. Key concepts in getting from data to insight include pattern discovery, quantification, statistical analysis, and visual communication.

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.

Data Analytics and Business Analytics

Variations of Purpose



Data Analytics and Business Analytics

Variations of Purpose

TWO VIEWS OF ANALYTICS

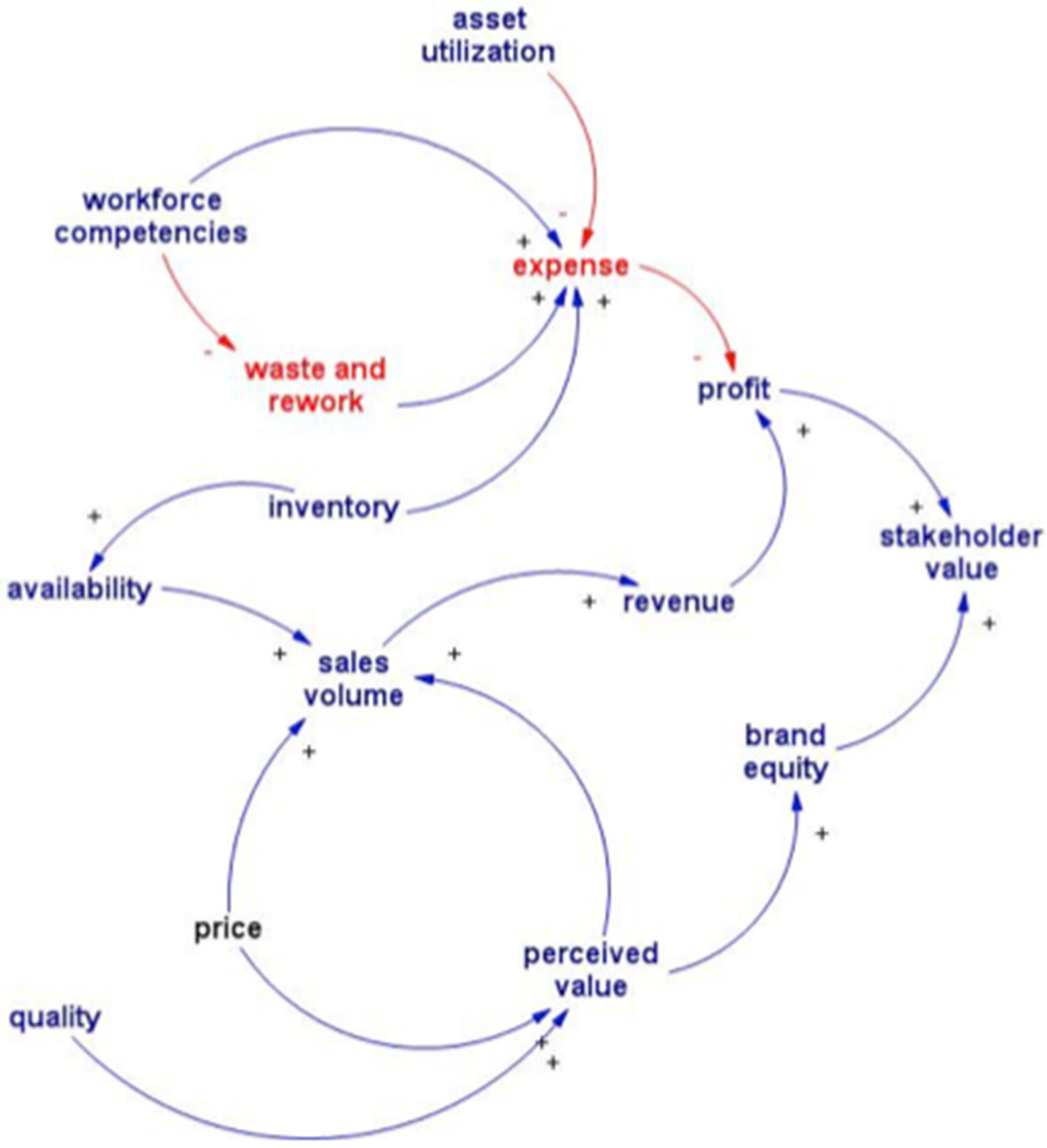
Analytics has two overlapping but distinctly different dimensions. Data analytics is performed by computers with the goal of finding patterns that build understanding and communicate insight. Business analytics applies understanding and insight together with business subject expertise. Business analytics is performed by people with the goal to find meaning that informs business decisions and actions.

Business analytics depends on data analytics to provide insights.

Data analytics depends on business analytics to have impact.

Why Analytics?

Cause and Effect



Why Analytics?

Cause and Effect

WHY THINGS HAPPEN

Knowing why things happen is important and valuable for analytic insight. Cause and effect is complex with every result driven by many influences. Profit, for example, is influenced by both revenue and expense. Further complexity is introduced with causal chains. The facing page illustrates several influences that form causal chains, such as:

- Expense influences profit. When expense increases profit decreases. When expense decreases profit increases.
- Asset utilization influences expense. When utilization increases expense decreases. When utilization decreases expense increases.
- Inventory costs influence expense. Larger inventory has higher expense. Smaller inventory has lower expense.
- Waste and rework influence expense. High waste brings high expense. Low waste reduces expense.
- Workforce competencies influence waste and rework – higher competency, less waste … lower competency, more waste.
- Workforce competencies influence expense – higher competency is more expensive than lower competency.

With this view it is clear that increasing profit isn't as simple as increase revenue and reduce expense. Decisions need to be made about where and how to reduce expense. Decision makers need to be aware of potential side-effects of actions. Cutting expense by lowering workforce competencies (eliminating training, for example) may have a high cost in waste and rework – perhaps even greater cost than the savings achieved by slashing the training budget.

CAUSAL MODELS AND ANALYTICS

Causal analysis has two distinct connections with analytics:

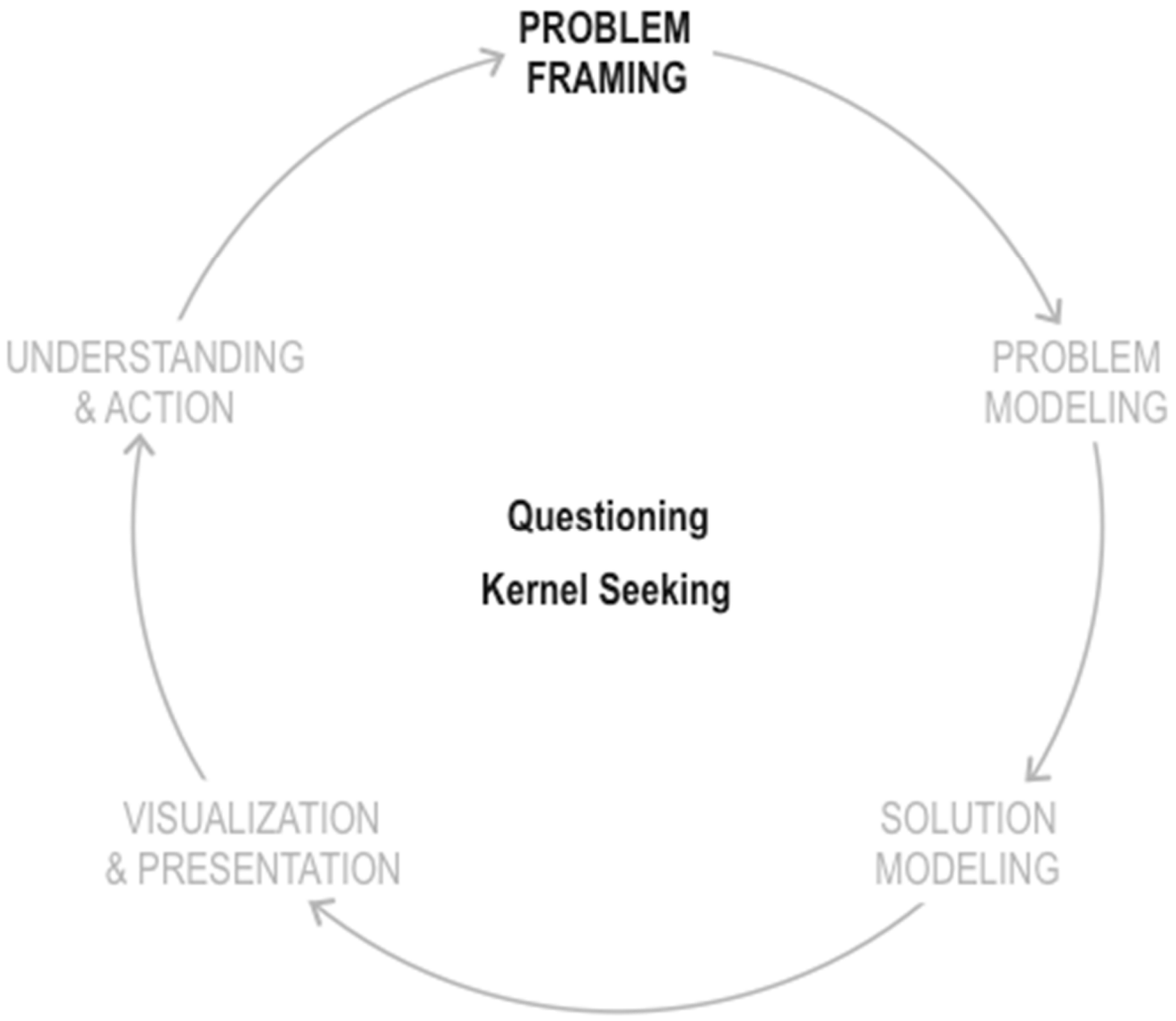
- Correlation is a frequently found pattern in analytics. Although correlation isn't always causal, the patterns are helpful to understand cause and effect and to construct causal models.
- Cause and effect in business is sometimes treated as "tribal knowledge" with decisions made based upon what we believe to be true. Whether formalized as a causal model or simply practiced as a mental model, the beliefs aren't always true. Data and analytics help to validate or refute beliefs about cause and effect.

WHY IT MATTERS

Causal models help managers find leverage points for change and avoid unintended consequences of side effects. They play an important role in getting from knowing why to deciding *what next.*

Analytics Processes

Problem Framing



Analytics Processes

Problem Framing

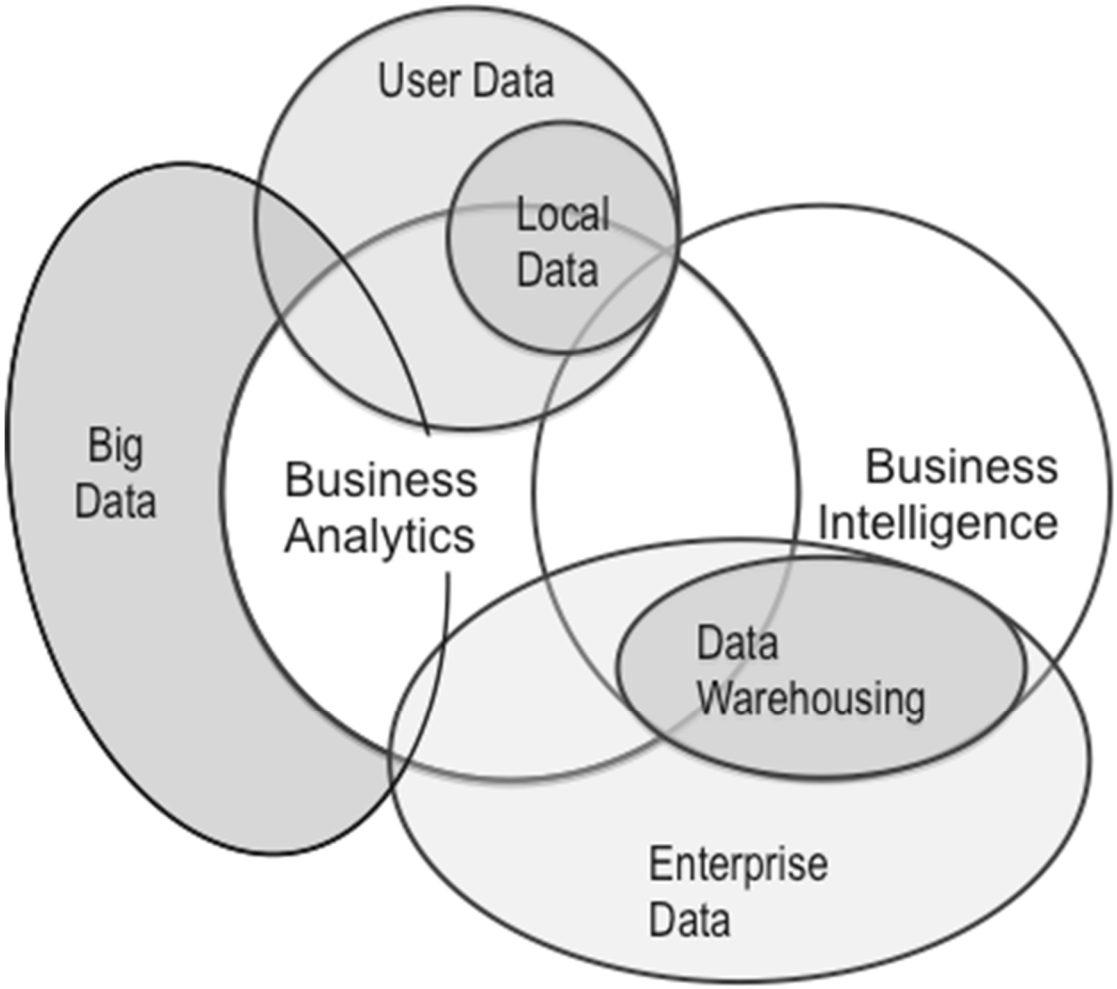
FROM VAGUE TO SPECIFIC

Most analytic problems begin with some uncertainty about the nature of the problem. Framing is the means to remove vagueness by adding specifics to our understanding of the problem. Problem framing is an important first step to avoid the "blind alleys" and unnecessary costs of data analysis with clear and well-defined purpose. Two techniques – questioning and kernel seeking – are useful for problem framing and can be used independently or together.

Well discuss these techniques in *Module 4 – Analytic Modeling.*

Analytics Foundations

Data



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

Analytics Foundations

Data

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.

Analytic data comes from many sources. Unlike BI, which primarily focuses on enterprise data and the data warehouse, the scope of analytic 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 analytic processes.
- Big data offers a variety of data sources to enrich the analytic 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

Analytics Stakeholders

The Participants

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.

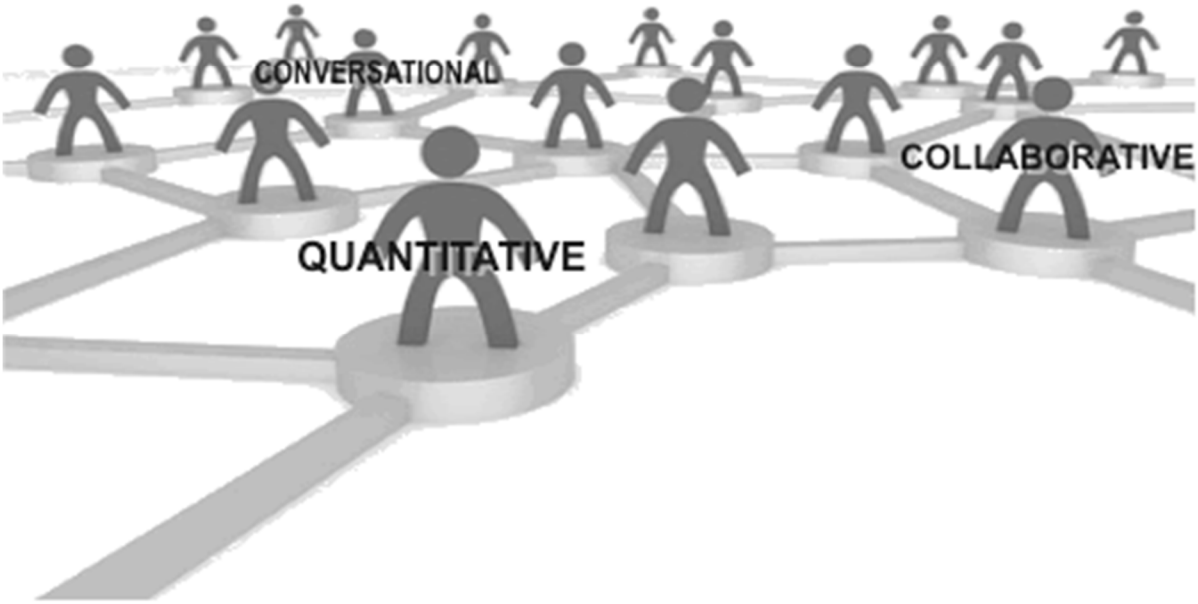
- Business analysts are the people who analyze business behaviors. They are responsible to evaluate and interpret the results of analytic models – to develop conclusions, identify alternatives, and make recommendations. In many organizations there is a high level of overlap between business managers 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 to understand analytic needs or problems, obtain and prepare data, and apply 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 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.

Analytics Culture

Values, Beliefs, and Competencies



Analytics Culture

Values, Beliefs, and Competencies

NUMERACY

Numeracy complements literacy and is sometimes called mathematical literacy. It is the ability to reason and to apply simple numerical concepts. Basic numeracy skills consist of comprehending fundamental arithmetics like addition, subtraction, multiplication, and division. Numeracy in business and analytics is fundamental to the ability to think quantitatively.

COLLABORATION

With multiple stakeholders – managers, analysts, modelers, and data experts – it is clear that business analytics is not an individual or solitary process. But the most effective analytics go well beyond teamwork to embrace collaborative analytics. Collaboration is the act of working jointly – two or more people combining their efforts toward achieving shared or intersecting goals. Collaborative analytics, then, is a set of analytic processes where the analysts work jointly and cooperatively to achieve shared or intersecting goals. Collaborative analytics includes data sharing, collective analysis and coordinated decisions and actions. The goals of conventional analytics are to find answers and make decisions. Collaborative analytics encompasses these goals but seeks to achieve more – to increase visibility of important business facts and to improve alignment of decisions and actions across the entire business.

CONVERSATION

The best analytics lead to conversations among people – discussion of what the data means, why things occur in business, the importance of predictions, and the best actions to take to drive positive business outcomes. Conversation is an important part of informed decision making.

DECISION STYLES

Understanding the nature of decisions and the decision style appropriate to each is a key competency when putting analytics into action. Decision making may be any of:

- Autonomous – a single individual has responsibility and authority to make the decision and may be informed by analytics.
- Advised – the decision maker is informed both by analytics and by conversation, interpretation, and advice of others.
- Influenced – the decision maker is informed by analytics and is influenced by interpretation and advice of others.
- Participative – the decision is make collectively by stakeholders who use analytics and conversation to arrive at a consensus decision.

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

Analytics Organizations

Organization Models

SELF SERVICE

The self-service model creates an environment where business units meet their own analytic needs with support of business-oriented tools, , architectures, frameworks, guidelines, examples, templates, etc. This model 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.

Analytics Capabilities

Business Capabilities

PLAN

EXECUTE

ADAPT

INNOVATE

Analytics Capabilities

Business Capabilities

ENABLING THE BUSINESS

The ultimate purpose of analytics is to enable business capabilities. At a high level the essential business capabilities are a continuum of planning, execution, adaptation, and innovation.

PLANNING

Planning is a process of thinking about and organizing the activities required to achieve a desired goal. Analytics support and inform planning in many ways that include:

- Knowing and quantifying the current state of something
- Understanding cause-effect relationships and identifying leverage points to make changes
- Forecasting and predicting future conditions with planning implications.

EXECUTING

Execution is the work of carrying out the activities described by a plan. Analytics support monitoring, tracking, and ability to detect both desired and unexpected outcomes of execution.

ADAPTING

Adaptation adjusts plans and activities to respond to unforeseen circumstances, unexpected outcomes, unintended side-effects, and other situations where plans don't align with realities.

INNOVATING

Innovation is the introduction of new products, processes, and methods to create new value opportunities. Product innovation may include any or all of new products for existing markets, new markets for existing products, and new products in new markets. Value, of course, derives from market growth and corresponding revenue growth. Innovation of processes and methods seeks new ways to do things, creating value through increased efficiency, greater effectiveness, higher quality, etc.

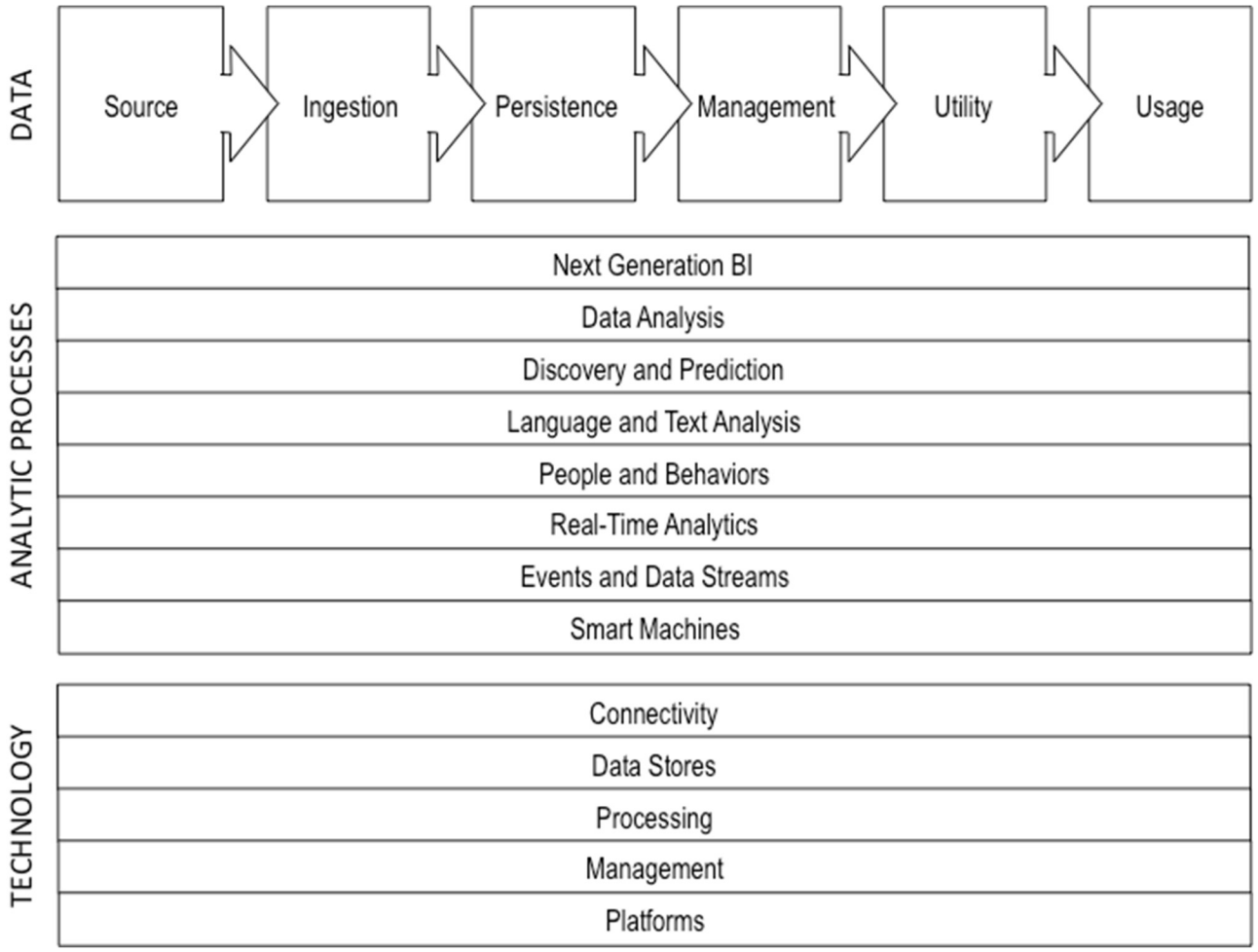
Transforming Data With Intelligence™

Module 3

Analytics Architecture

The Big Picture

Architecture Overview



The Big Picture

Architecture Overview

CONTEXT

The implementation of analytics is complex, with many different components to working together to create positive outcomes and real business impact. A team of stakeholders with diverse backgrounds is needed to realize the potential of analytics. Some stakeholders create data centric applications and mathematical models. Other stakeholders consume information from applications and models to derive new insights. Ultimately key business stakeholders will drive business impact from the insights.

PURPOSE

Architecture for analytics defines the business objectives and success criteria of the "analytics system," the necessary components to achieve the objectives, and the connections and interactions among the components.

The architecture provides a reference model to help all stakeholders understand critical dependencies and structure of the analytics system. It enables alignment across diverse stakeholder groups to support planning, implementation and operational decisions needed to achieve the overall vision.

COMPONENT CATEGORIES

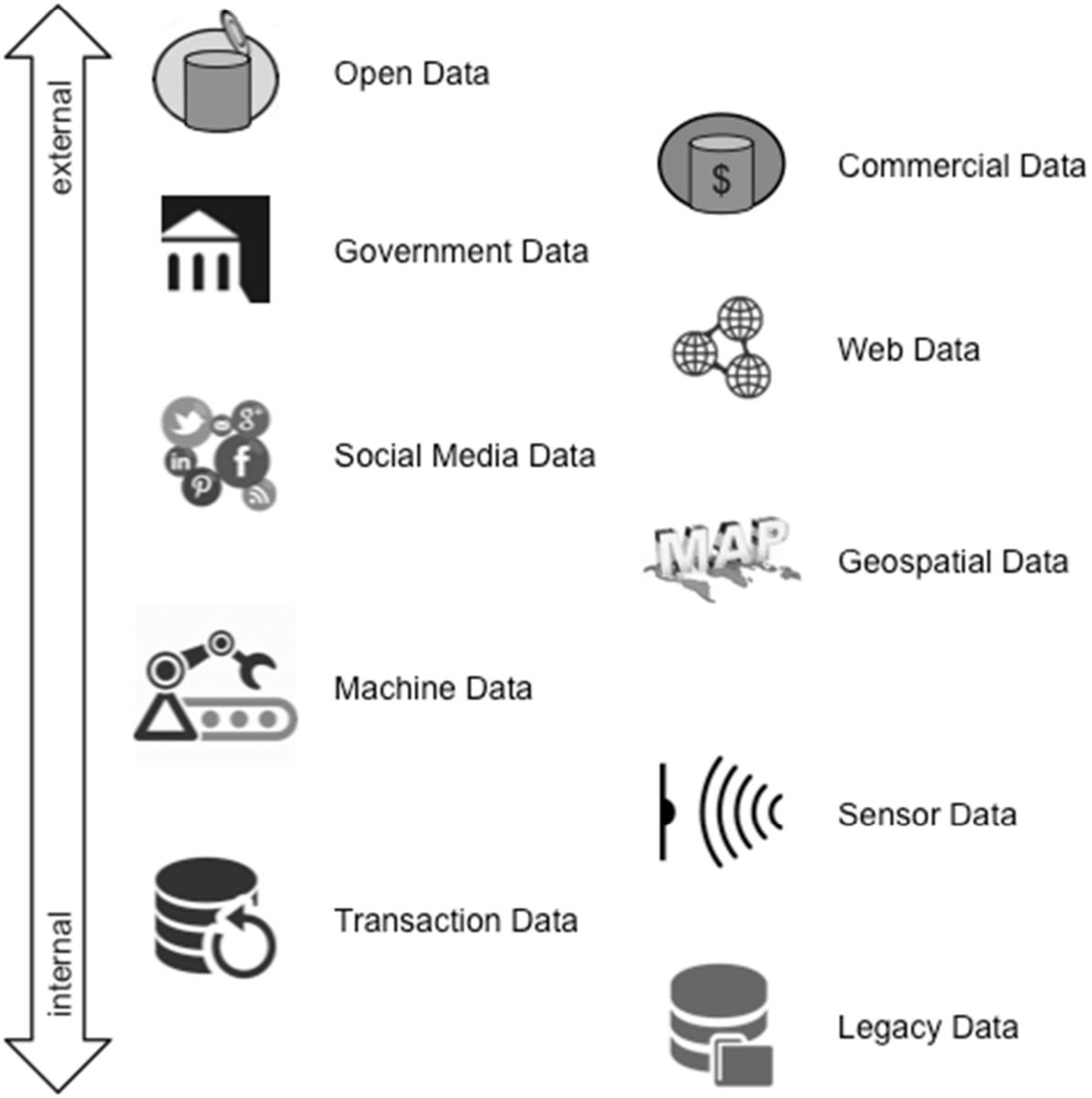
The major categories of components required for analytics success are listed below.

- Data
- Process
- Technology

These component categories are described in greater detail through the remainder of this course.

Data Architecture

Data Sources and Types



Data Architecture

Data Sources and Types

CONTEXT

The available data for analysis is diverse and rich with opportunities. As an architectural component, source data describes the available data in terms of origination, context, structure and latency.

CATEGORIES OF DATA SOURCES

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 – licenced for free re-use
- 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

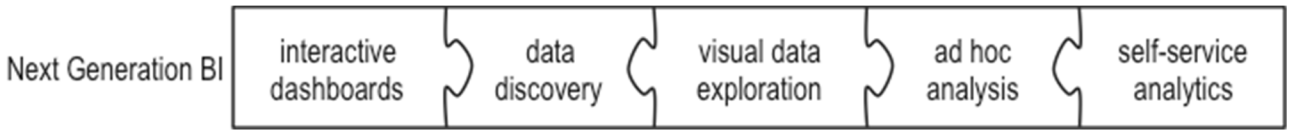
Data Latency

- 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.

Process Architecture

Next Generation BI



Process Architecture

Next Generation BI

CONTEXT

Process architecture describes key functionality needed to enable the analytic and business capabilities delivered by the BI systems.

ADVANCES IN BUSINESS INTELLIGENCE

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

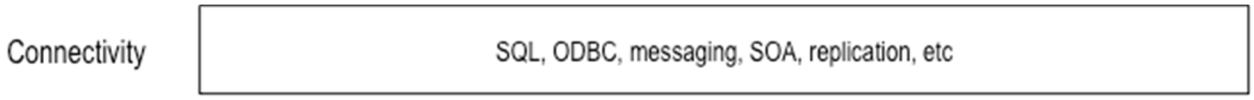
BI is now recognized to be an active system that delivers functionality in addition to information. A variety of new innovations enable these advances by transforming BI from a passive to a functional system.

The following items are examples of advances in BI that enable segments 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 analyse 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.

Technology Architecture

Connectivity



Technology Architecture

Connectivity Architecture

OVERVIEW

The Technology Architecture segment of the Analytics Architecture defines the hardware, software, communications, algorithms and storage components from a capability perspective.

The first layer in this segment is the Connectivity component.

CONNECTIVITY

The Connectivity component defines the required capabilities for methods and protocols to establish physical connections to all necessary source data components. Examples are:

- SQL
- Messaging
- Services
- Replication
- Virtualization

The data layer of the analytics architecture provides the requirements for connectivity.

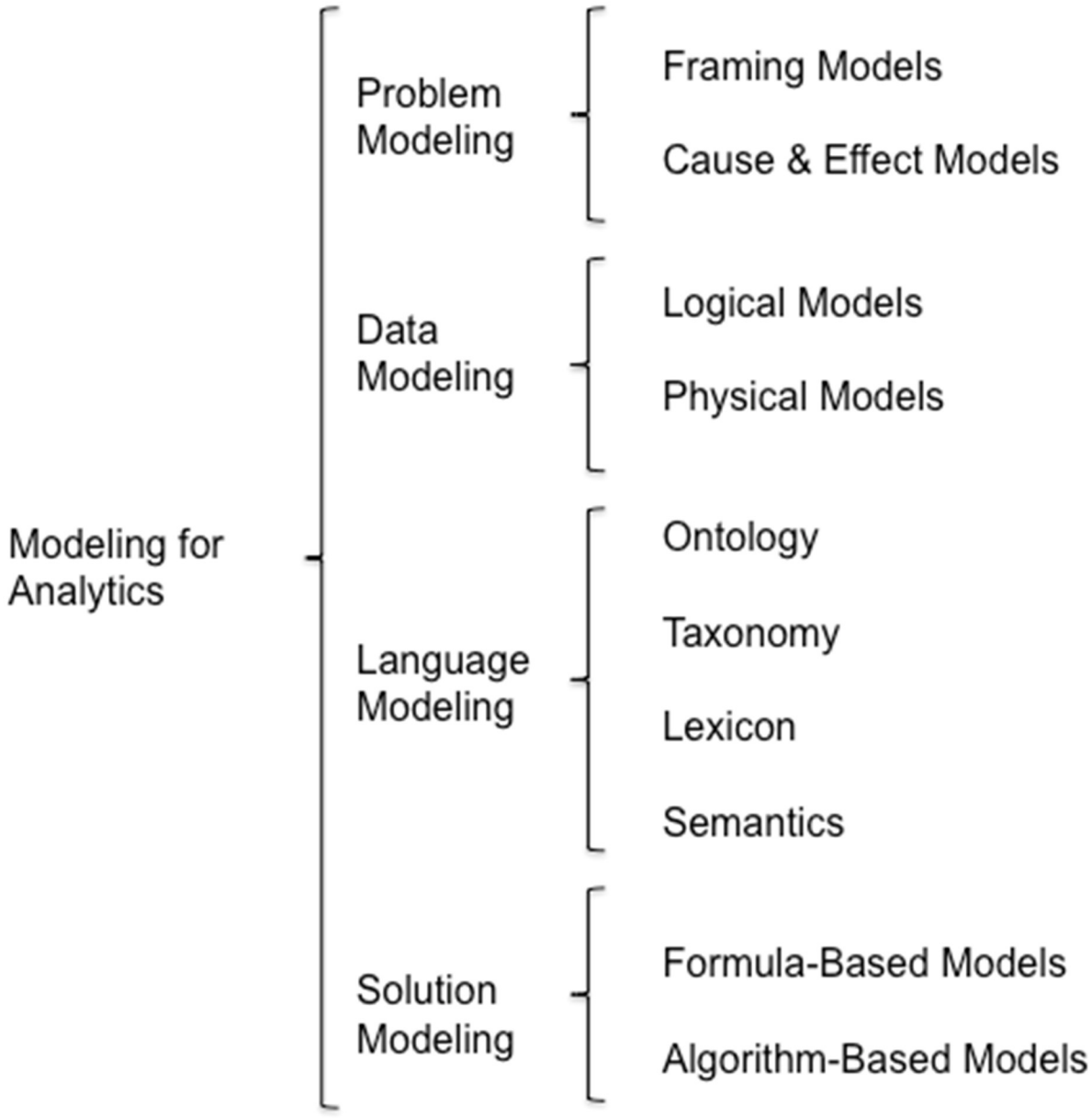
Transforming Data With Intelligence™

Module 4

Analytic Modeling

The Roles of Models

Why Model?



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 analytic 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 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.

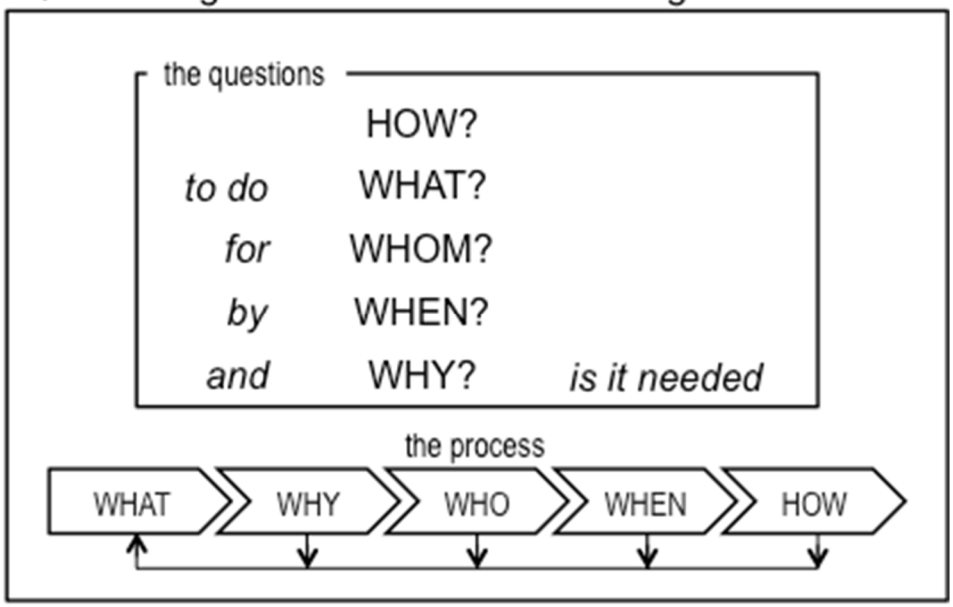
LANGUAGE MODELING When data for analytics includes text, analysis processes must parse and inspect text to turn it into useful and quantitative data. Text analysis uses four kinds of models: Ontological models are a linguistic structure to describe things in the real world and how they are related. Taxonomic 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 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.

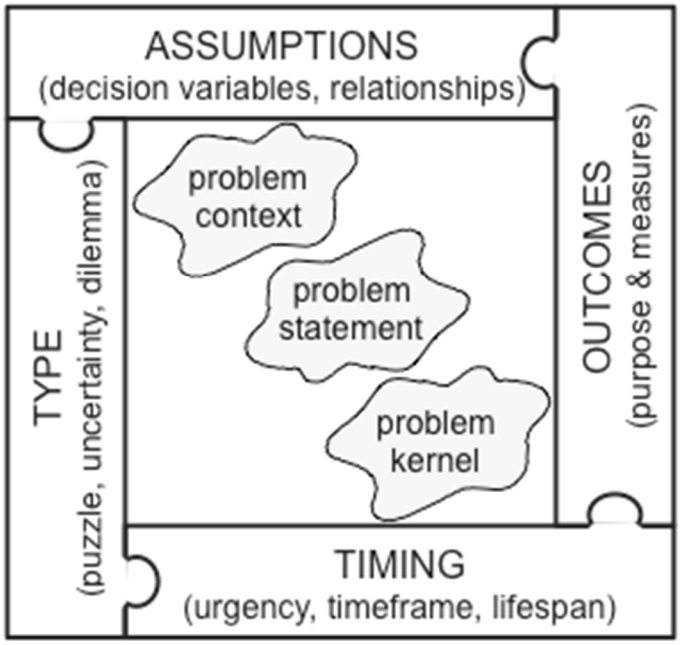
Kinds of Models

Framing Models

Questioning Model for Problem Framing



Kernel Seeking Model for Problem Framing



Kinds of Models

Framing Models

UNDERSTANDING THE PROBLEM

The facing page illustrates two kinds of framing models – questioning and kernel seeking.

The questioning model asks

How to do *what* by *when* for *whom*, and *why* is it needed?

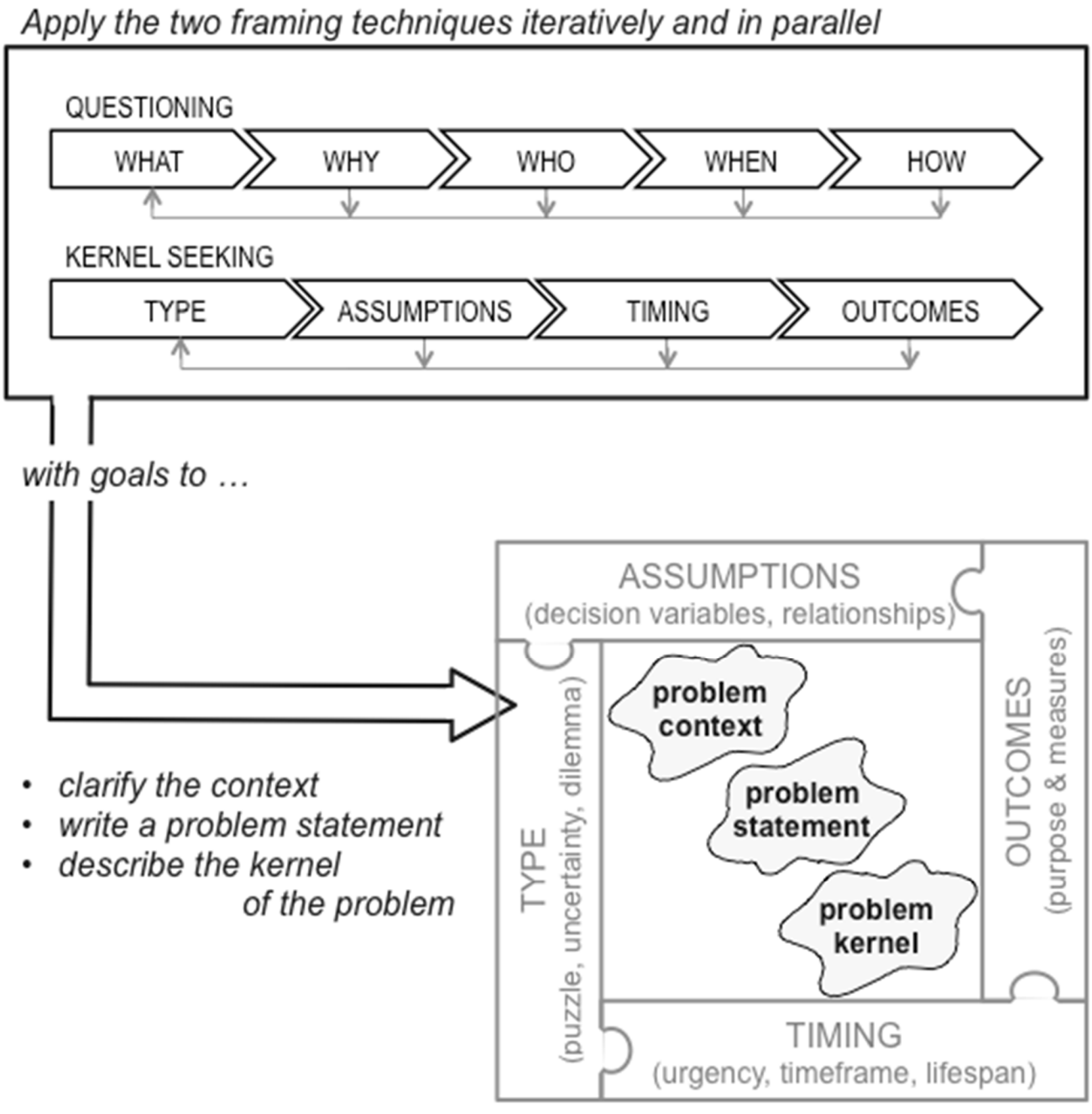
The sequence in which questions are addressed, however, is likely to vary from the sequence expressed above. It is common to ask: What needs to be done? Why is it needed? Who needs it? By when? How to make it happen?

The kernel seeking model works from context, through problem statement, to express the essence of the problem as a single simple sentence.

You'll have opportunity to work with both types of models and to see how they work together as part of a problem framing activity.

Problem Modeling

Framing the Problem



Problem Modeling

Framing the Problem

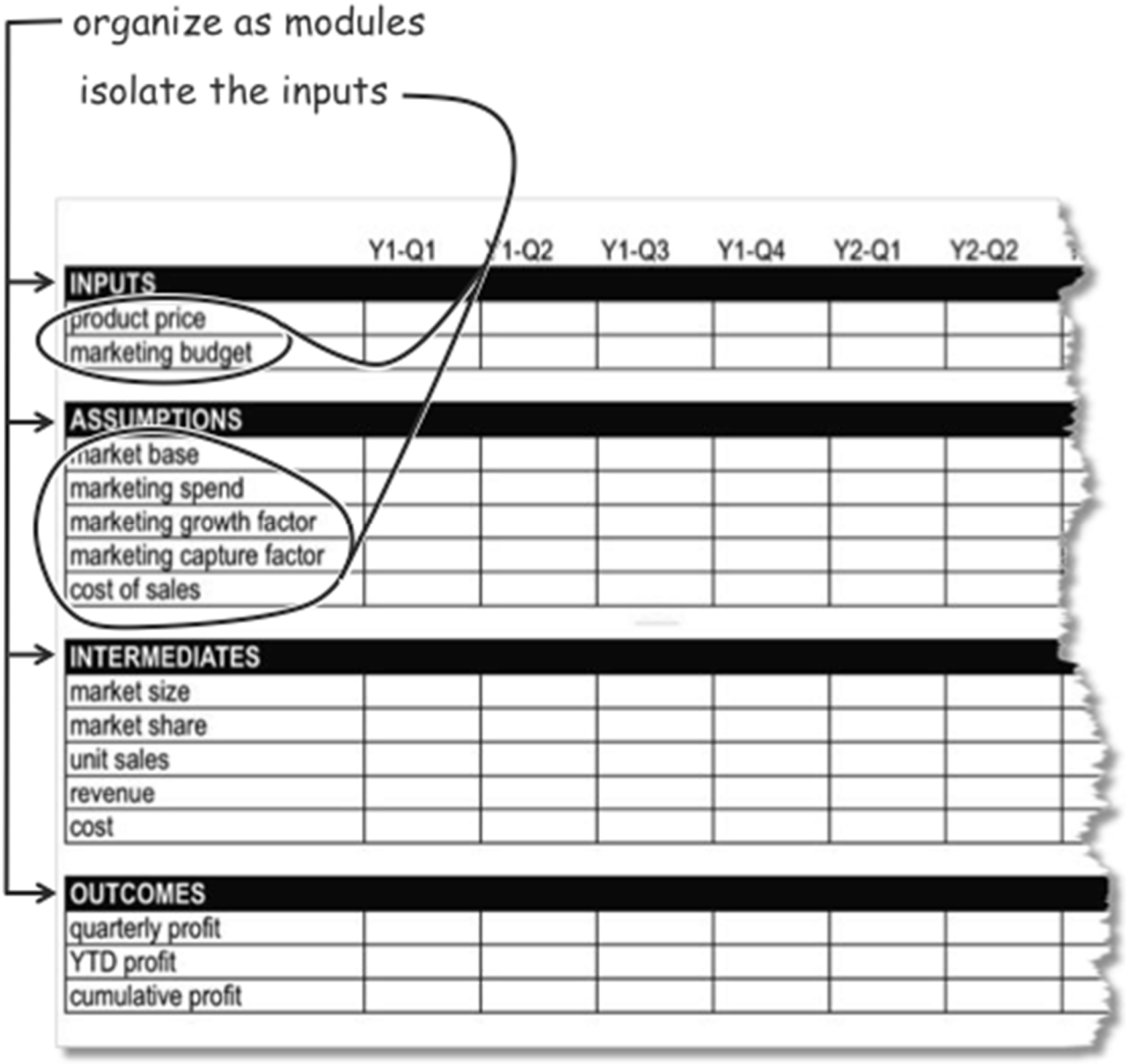
USING TWO TECHNIQUES

Problem framing is a necessary but often elusive first step in analysis. Clearly stating the analysis problem can be difficult, so this step is sometimes bypassed to "get on with real modeling." Yet modeling for an ill-defined problem with vague expectations of results is a poor approach.

Combine the two techniques – questioning and kernel seeking – to help overcome the challenges of problem framing.

Solution Modeling

Formula Based Modeling – Structuring



Solution Modeling

Formula Based Modeling – Structuring

SPREADSHEET ENGINEERING

The technique of "spreadsheet engineering" is used in this course to illustrate many techniques of modeling analytic solutions. It is not recommended, nor is it practical, to meet all of your analytic 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 analytic 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

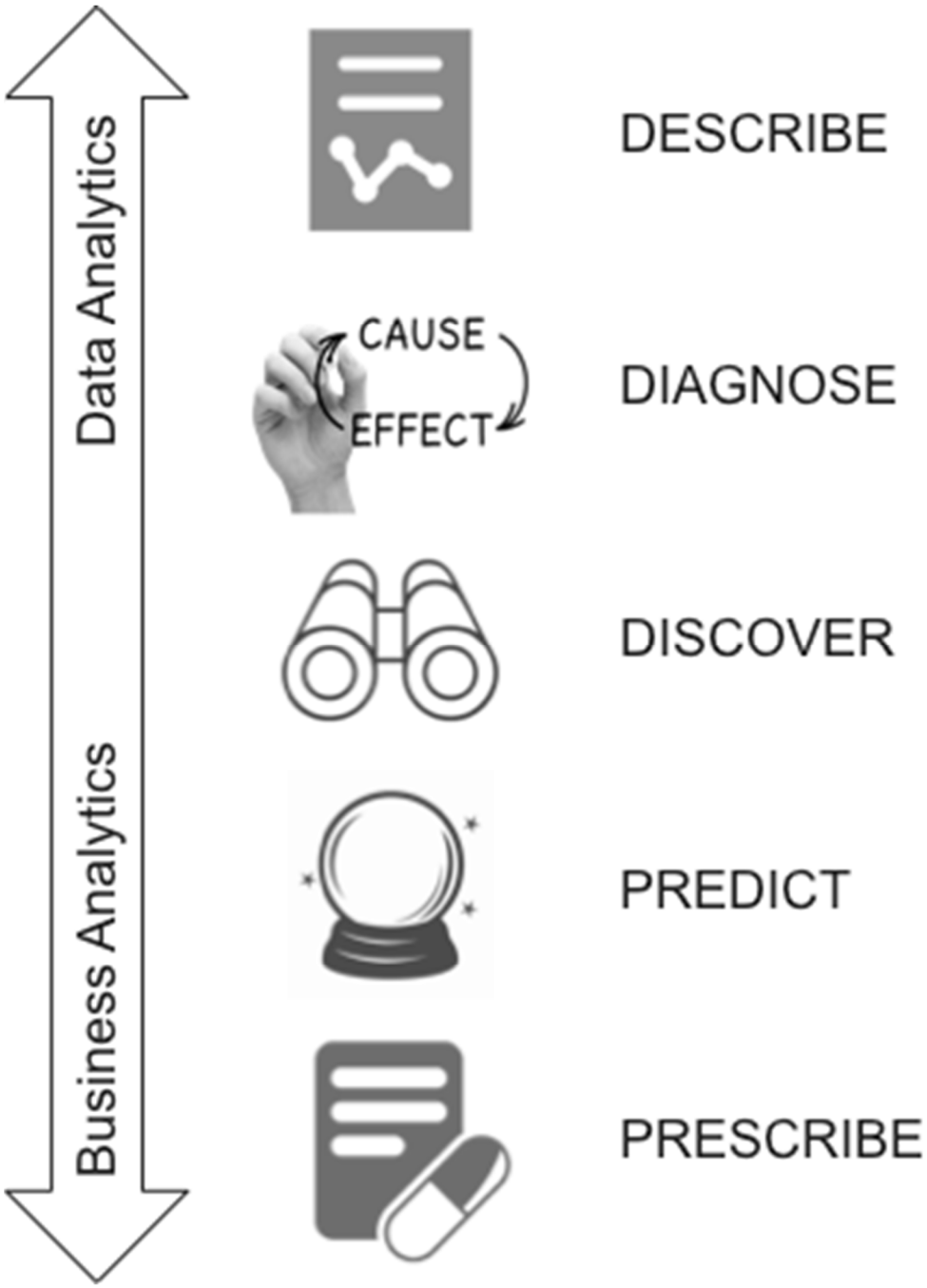
tdwi
Transforming Data With Intelligence™

Module 5

Applied Analytics

Five Kinds of Analytics

What We Do



Five Kinds of Analytics

What We Do

DIFFERENCES IN ANALYTICS FOCUS

Analytics is not a "one size fits all" endeavour. 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 build 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.

