Transforming Data
With Intelligence™

TDWI Data Quality Management

Techniques for Data Profiling, Assessment, and Improvement

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.

TABLE OF CONTENTS

COURSE OBJECTIVES

To learn:

- ✓ ***Techniques for column, table, and cross-table data profiling***
- ✓ ***How to analyze data profiles and find the stories within them***
- ✓ ***Subjective and objective methods to assess and measure data quality***
- ✓ ***How to apply OLAP and performance scorecards for data quality management***
- ✓ ***How to get beyond symptoms and understand the real causes of data quality defects***
- ✓ ***Data cleansing techniques to effectively remediate existing data quality deficiencies***
- ✓ ***Process improvement methods to eliminate root causes and prevent future defects***

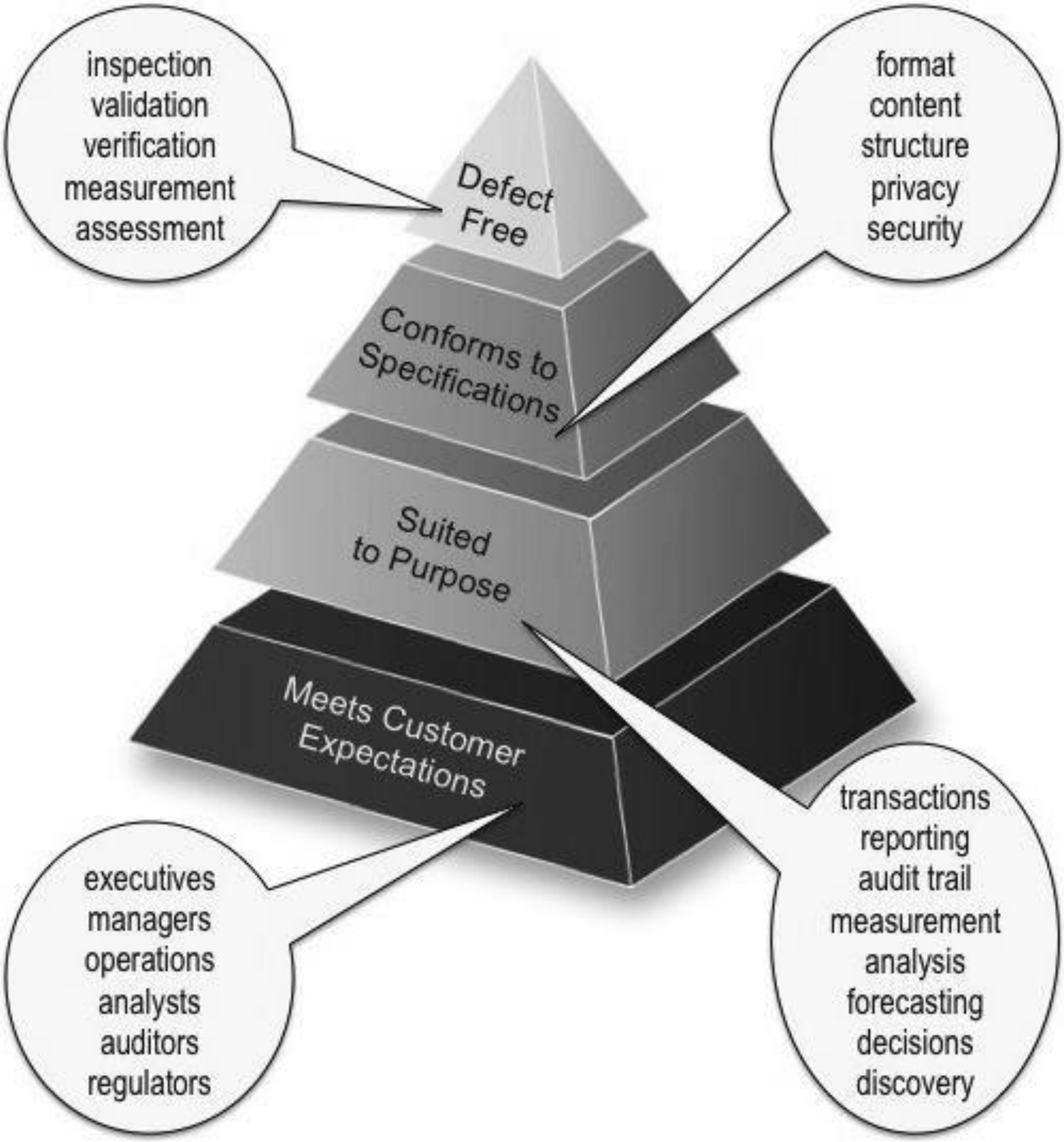
Transforming Data With Intelligence™

Module 1

Data Quality Basics

Data Quality Concepts

Defining Data Quality



Data Quality Concepts

Defining Data Quality

QUALITY DEFINITIONS

Merriam-Webster dictionary defines quality as “degree of excellence.” The important point here is that quality is not an absolute, but something that exists in degrees. One common definition describes high quality as **defect free**. This interpretation comes from the community of quality practitioners who base their practice on the principle of zero defects. They define quality as **conformance to specifications** and defects as variance from specifications. Another widely used definition states that quality is **suitability to purpose** – a thing is of high quality when it is well suited to the purpose that is its intended use, and it is of poor quality when badly suited to its purpose. The principles of Total Quality Management (TQM) define quality as consistently **meeting customer expectations**. This principle promotes the idea that quality doesn’t reside within a product; it can only be judged in relation to the expectations of the customer using the product.

DATA AND DEFECTS

Defect-free data requires identification of the things that are data defects (more about this later), after which you can manage by inspecting data to find defects, by validating and verifying data as free of defects, and by measuring defects as part of data quality assessment.

DATA AND SPECIFICATIONS

Conformance to specifications requires formal data specifications, which may address any or all of data format, content, and structure as well as usage-oriented specifications such as those for data privacy and security. Data quality management will test data against specifications.

DATA AND PURPOSE

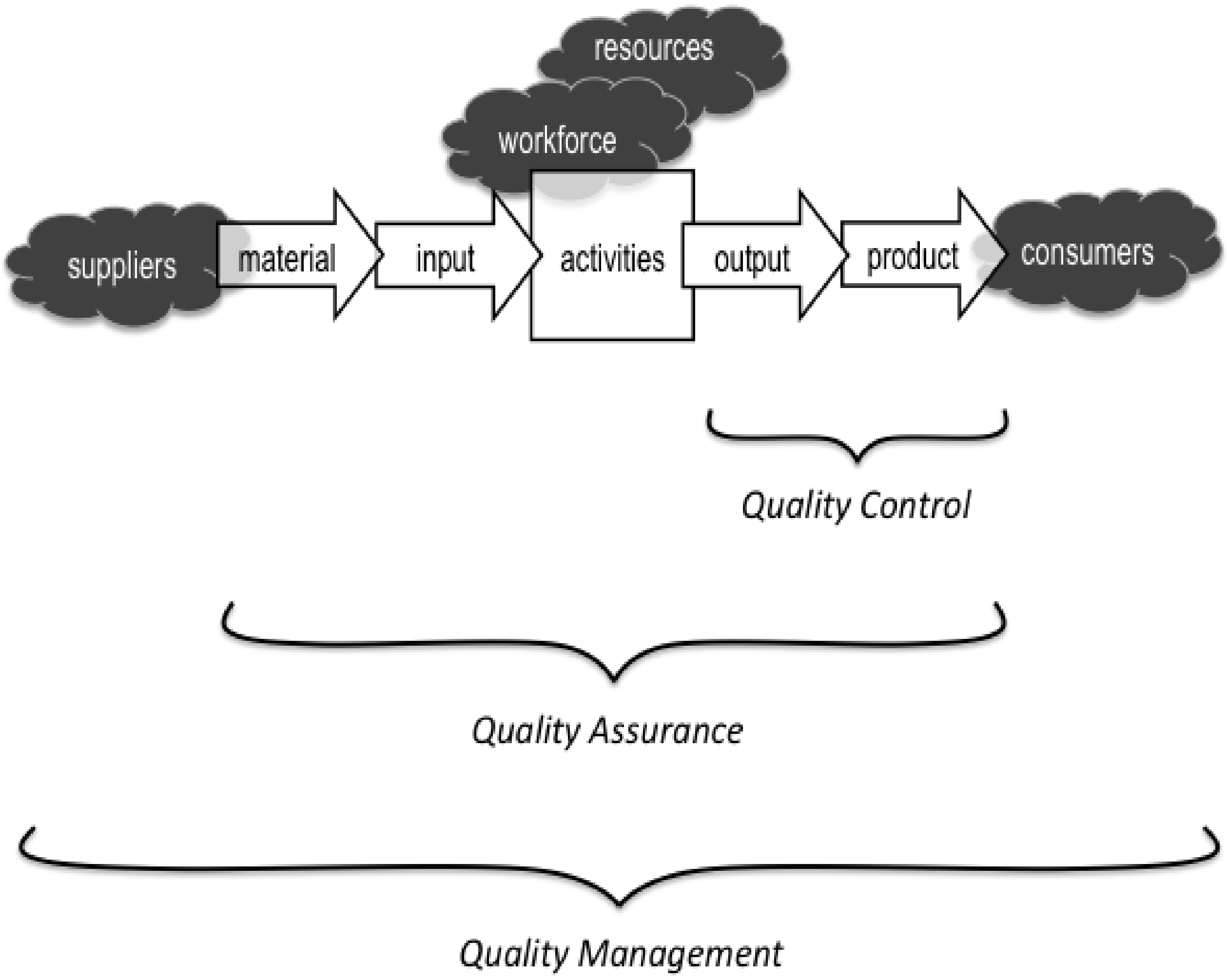
Suitability to purpose must consider all purposes for which data is used, ranging from business transactions and operational reporting to business intelligence and analytics. Expect the quality criteria to vary widely among the different uses. Variations in quality criteria increase the level of difficulty in data quality management, but attention to them makes quality management efforts more effective and far-reaching.

DATA AND EXPECTATIONS

Data quality as meeting customer expectations must consider the wide range of data and information consumers. Expect wide variation in the expectations through the range of consumers, both internal and external. Quality management implications of varied expectations are much like those for varied purpose – greater complexity and greater impact.

Data Quality Processes

Quality Control, Assurance, and Management



Data Quality Processes

Quality Control, Assurance, and Management

SCOPE OF QM

Comprehensive quality management focuses on process as well as product, and on things external to the process as well as process internals.

Every product is the result of a process – a set of activities that receive raw material and create the product through value-adding steps. External to the process are suppliers of material, consumers of products, and the workforce and resources to perform the activities. This construct is as true for data as for any other product.

LEVELS OF QM

Quality management can be performed at each of three levels:

- Quality control (QC) is the narrowest view of QM, and is based on checking the product for defects before it is released.

- Quality assurance (QA) broadens the view by looking “up the line” to check quality at the activities and materials stages of production. QA includes QC and more.

- The end-to-end view of quality management (QM) looks outside as well as inside the production process. QM extends quality practices to include external factors of suppliers, workforce, resources, and consumers (customers). End-to-end QM fits well with the definition of quality as meeting customer expectations. QM includes both QA and QC, but it expands to include quality planning and quality improvement.

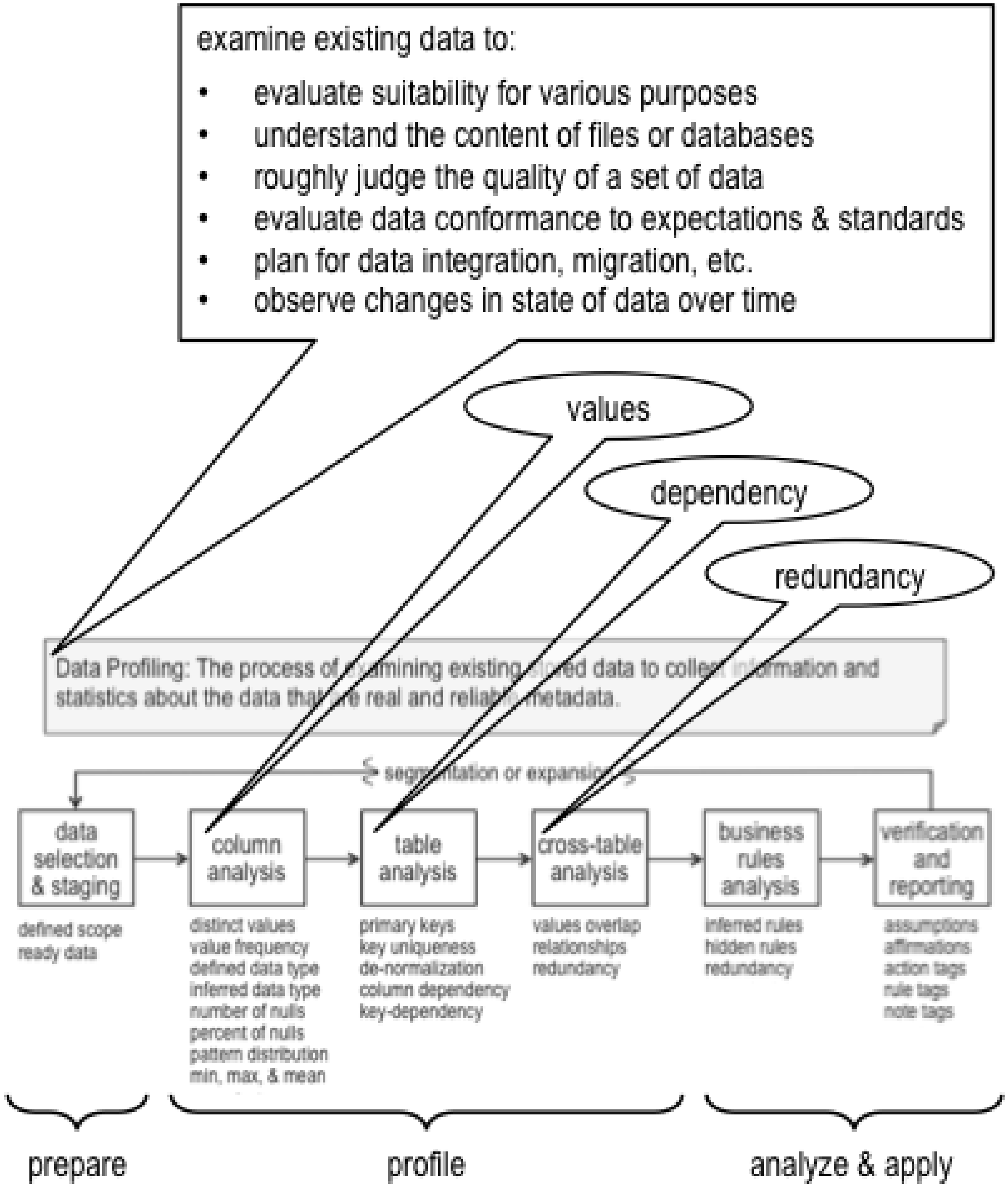
Transforming Data With Intelligence™

Module 2

Profiling Data

Data Profiling Concepts

Purpose and Processes



examine existing data to:

- evaluate suitability for various purposes
- understand the content of files or databases
- roughly judge the quality of a set of data
- evaluate data conformance to expectations & standards
- plan for data integration, migration, etc.
- observe changes in state of data over time

Data Profiling: The process of examining existing stored data to collect information and statistics about the data that are real and reliable metadata.

Data Profiling Concepts

Purpose and Processes

WHY PROFILE?

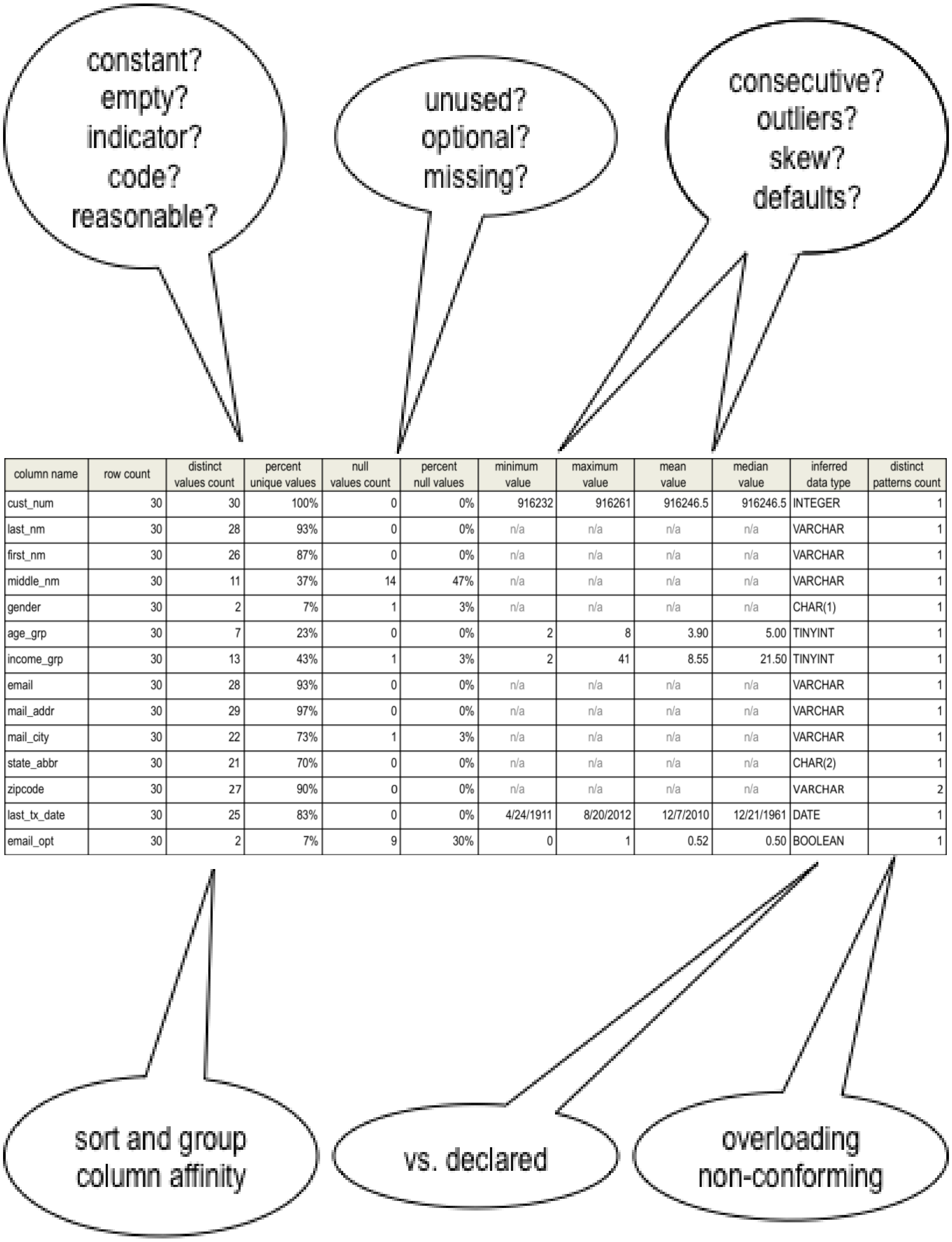
Data profiling is the work of understanding the data by looking at the data. While looking at the data may seem an obvious necessity to some, it is often overlooked. The tendency to review data models, descriptions, definitions, and program code causes many to overlook the obvious. And those who do look at the data often do so in an unstructured way that leads to seeing only that which is expected.

STAGES AND STEPS

Data profiling overcomes the pitfalls of unstructured data review by systematically examining data to describe the realities found in the data. Data profiling is a process that involves three stages: preparation, building of data profiles, and analysis of those profiles. Building profiles includes three data analysis steps: column analysis, table analysis, and cross-table analysis.

Analyzing Data Profiles

Column Profiles



column name	row count	distinct values count	percent unique values	null values count	percent null values	minimum value	maximum value	mean value	median value	inferred data type	distinct patterns count
cust_num	30	30	100%	0	0%	916232	916261	916246.5	916246.5	INTEGER	1
last_nm	30	28	93%	0	0%	n/a	n/a	n/a	n/a	VARCHAR	1
first_nm	30	26	87%	0	0%	n/a	n/a	n/a	n/a	VARCHAR	1
middle_nm	30	11	37%	14	47%	n/a	n/a	n/a	n/a	VARCHAR	1
gender	30	2	7%	1	3%	n/a	n/a	n/a	n/a	CHAR(1)	1
age_grp	30	7	23%	0	0%	2	8	3.90	5.00	TINYINT	1
income_grp	30	13	43%	1	3%	2	41	8.55	21.50	TINYINT	1
email	30	28	93%	0	0%	n/a	n/a	n/a	n/a	VARCHAR	1
mail_addr	30	29	97%	0	0%	n/a	n/a	n/a	n/a	VARCHAR	1
mail_city	30	22	73%	1	3%	n/a	n/a	n/a	n/a	VARCHAR	1
state_abbr	30	21	70%	0	0%	n/a	n/a	n/a	n/a	CHAR(2)	1
zipcode	30	27	90%	0	0%	n/a	n/a	n/a	n/a	VARCHAR	2
last_tx_date	30	25	83%	0	0%	4/24/1911	8/20/2012	12/7/2010	12/21/1961	DATE	1
email_opt	30	2	7%	9	30%	0	1	0.52	0.50	BOOLEAN	1

Analyzing Data Profiles

Column Profiles

COLUMN ANALYSIS

The list of things that can be discovered through column analysis is long. Common column analysis discoveries include:

- Distinct values analysis finding
  - Constants – only one value that is not blank and not zero
  - Empty columns – only one value that is either blank or zero
  - Indicators – number of distinct values exactly 2 (y/n, t/f, or 0/1)
  - Codes – number of distinct values in single or low double digits
- Null values analysis finding
  - Unused columns – 100% null values
  - Optional columns – percent of null values is relatively high
  - Missing data – percent of null values is relatively low
- Value distribution analysis finding
  - Consecutive numbers –
    row count = maximum value – minimum value + 1
    (small variance may mean some missing numbers in a sequence – not important in some cases but what about check register?)
  - Outliers – exceptionally high or low values, useful to look at top-ten and bottom-ten lists
  - Skew – substantial difference between mean and median
  - Default – exceptionally high frequency of a single value
  - Ranges and clusters – apparent ranges, clusters or gaps
- Distinct patterns analysis finding
  - Overloaded columns – two or three distinct patterns
  - Non-conforming columns – many distinct patterns such as phone numbers

METADATA MATCHING

Beyond the basic profile analysis described above, compare the profiles with your knowledge and with other metadata that is available.

- Check valid values by comparing distinct values with reference tables
- Compare declared data type with inferred data type
- Column affinity – Sorting by distinct values count to group similar columns (i.e., zipcode_low and zipcode_high or billing_state and shipping_state)
- Column affinity – Sorting by distinct values count will often group columns of similar data (i.e., zipcode columns or state abbreviations)

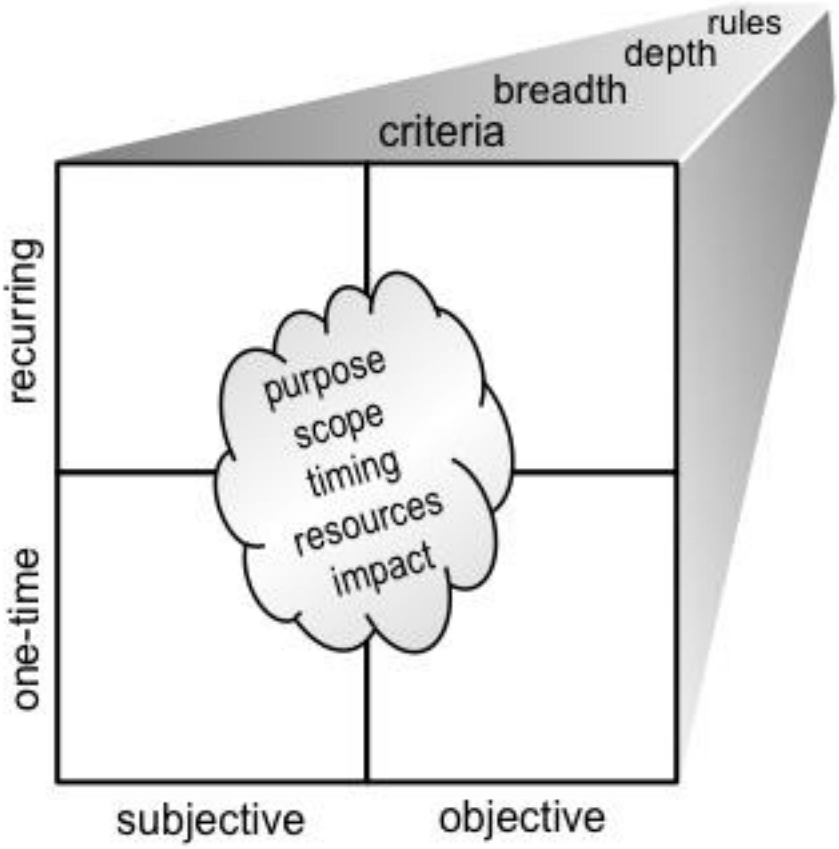
tdwi
Transforming Data
With Intelligence™

Module 3

Assessing Data Quality

DQ Assessment Concepts

DQ Assessment Defined



DQ Assessment Concepts

DQ Assessment Defined

DEFINITION

A multi-dimensional evaluation of the condition of data relative to any or all of the common definitions of quality:

- Defect free
- Conforming to specifications
- Suited to purpose
- Meeting customer expectations

DIMENSIONS AND VARIATIONS

Two types of assessment can be performed – subjective and objective. A subjective assessment measures perceptions and beliefs of people who work with data, and is best matched to quality definitions for purpose and expectations. Objective assessment is a better fit for the more tangible definitions for specifications and defects.

Assessment may be performed either as a one-time activity or as a recurring process. Ideally, every data quality management program includes continuous and ongoing assessments. One-time assessment is most appropriate to special circumstances such as assessing the source data for a data conversion project.

Specific criteria vary between objective and subjective assessment, and with the breadth and depth of assessment that is needed. Objective assessment extends beyond criteria to include data quality rules. The set of rules to be tested is directly related to breadth and depth of assessment.

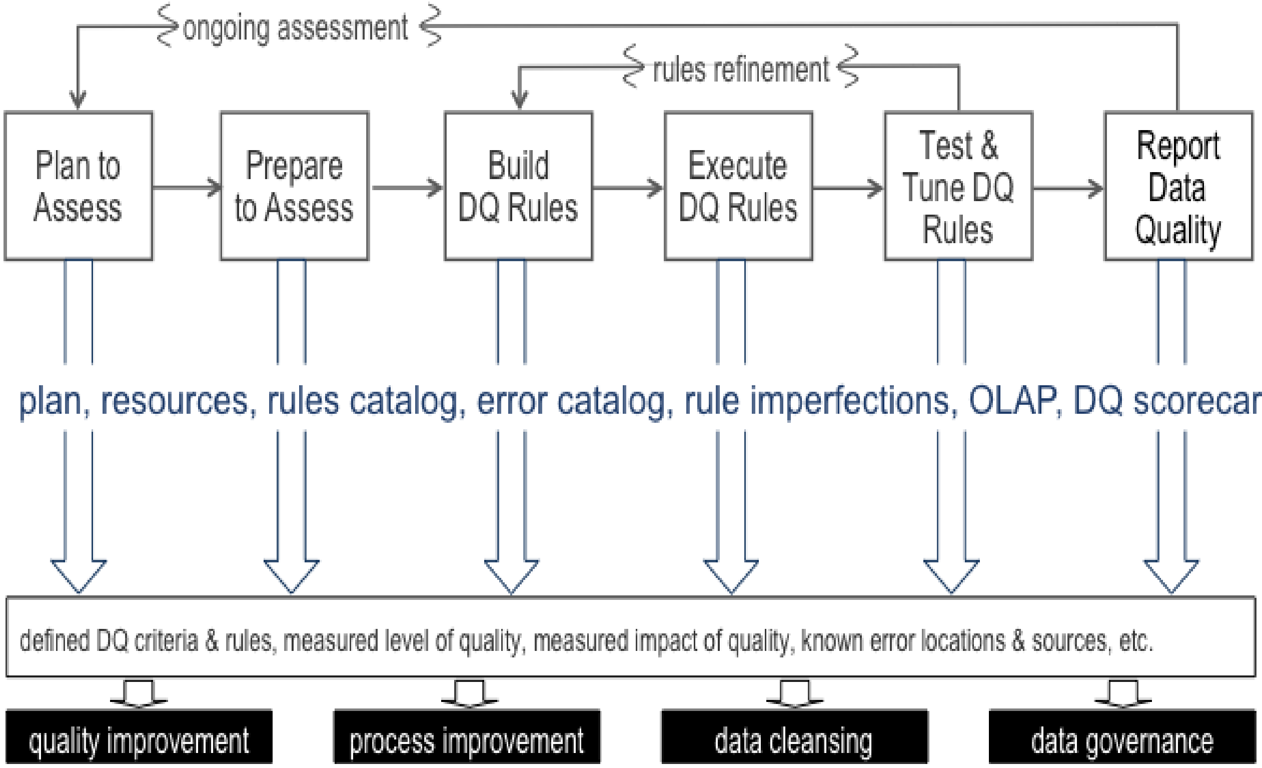
Choosing the type (or types) of assessment – one-time or recurring, subjective or objective – is guided by several factors including:

- Purpose of assessment
- The scope of data to be assessed
- Timing and time constraints
- Available resources
- Impact that you want to achieve
- Desired breadth and depth

With all of these variables in play it is expected that you'll need to perform many assessments in a DQ program. Becoming skilled at assessment is fundamental to DQ success.

Assessment in Practice

Assessment and Projects


ongoing assessment
rules refinement
Plan to Assess
Prepare to Assess
Build DQ Rules
Execute DQ Rules
Test & Tune DQ Rules
Report Data Quality
plan, resources, rules catalog, error catalog, rule imperfections, OLAP, DQ scorecard
defined DQ criteria & rules, measured level of quality, measured impact of quality, known error locations & sources, etc.
quality improvement
process improvement
data cleansing
data governance

Assessment in Practice

Assessment and Projects

ASSESSMENT AS PROJECTS

Each data quality assessment that you perform is a project that includes steps for planning, preparation, development, testing, execution, and delivery. All of the project management disciplines that are effective for other kinds of projects work equally well for DQ assessment.

ASSESSMENT IN SUPPORT OF PROJECTS

All of the common data quality management projects – data cleansing, process improvement, and quality improvement – begin with assessment. Only by assessing data quality can you know which data to cleanse, which processes to improve, or where to focus quality improvement efforts.

Although not a project but an ongoing program, data governance activities also benefit from data quality assessment. Effective governance requires feedback. For a quality-focused data governance program, assessment produces the feedback that is needed.

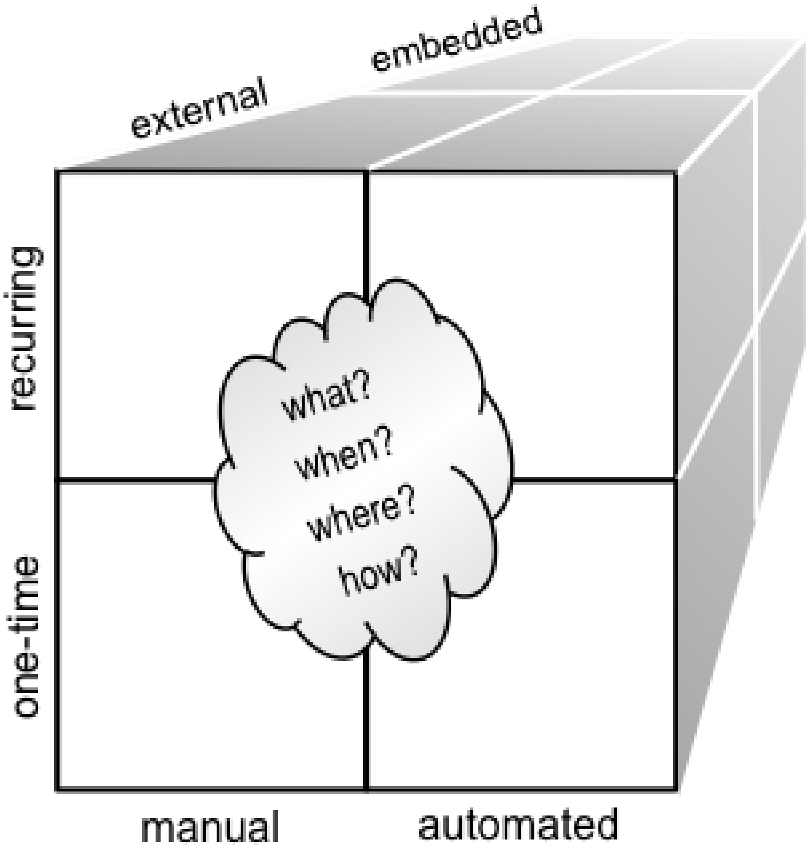
Transforming Data With Intelligence™

Module 4

Fixing Data Quality Defects

Data Cleansing Concepts

Data Cleansing Defined



Data Cleansing Concepts

Data Cleansing Defined

DEFINITION

Data cleansing is the act of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database. It is a process of finding and removing data quality defects. Cleansing may involve removing defective data from the collection, obtaining correct data from an alternate source, or adjusting defective data to comply with data quality rules.

DIMENSIONS AND VARIATIONS

Data cleansing may be:

- manual (performed by people) or automated (performed by computer)
- one-time (a single-instance repair) or recurring (regular or periodic processing)
- embedded (integrated into existing processes) or external (performed as a stand-alone process).

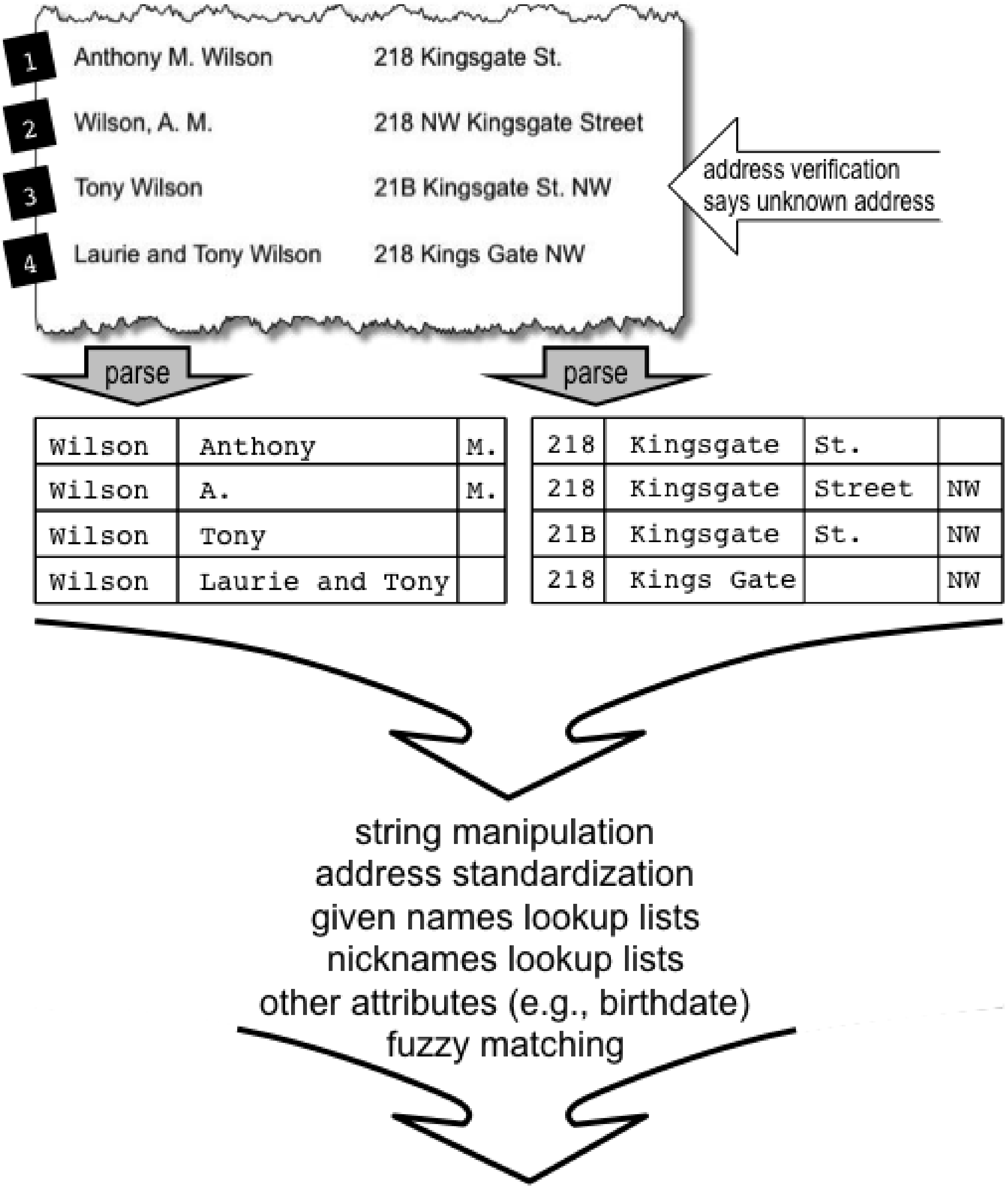
These options combine in some interesting ways – embedded, automated, recurring for example; or external, manual, one-time. A complete data cleansing solution typically uses a mix-and-match approach with several options.

High-level questions for each cleansing activity include:

- What to cleanse – which data and which defects?
- When to cleanse – at what point in business and systems schedules?
- Where to cleanse – at what point in the flow of data and processes?
- How to cleanse – using what methods and workflow?

Procedural Data Cleansing

Names and Addresses



1	Anthony M. Wilson	218 Kingsgate St.
2	Wilson, A. M.	218 NW Kingsgate Street
3	Tony Wilson	21B Kingsgate St. NW
4	Laurie and Tony Wilson	218 Kings Gate NW

Wilson	Anthony	M.
Wilson	A.	M.
Wilson	Tony	
Wilson	Laurie and Tony	

218	Kingsgate	St.	
218	Kingsgate	Street	NW
21B	Kingsgate	St.	NW
218	Kings Gate		NW

string manipulation
address standardization
given names lookup lists
nicknames lookup lists
other attributes (e.g., birthdate)
fuzzy matching

Records 1, 2, and 3 are high probability match. Record 4 contains a match, but it also contains a 2nd entity (person) and may require split into two records.

Procedural Data Cleansing

Names and Addresses

FINDING REDUNDANCY

Matching applies procedures to find things that appear to be identical. This is a key step in recognizing redundancy and an essential part of automated de-duplication.

Matching people, for example, on the basis of name and address is relatively easy when names and addresses are standardized. This may imply some standardization and perhaps some parsing or string manipulation as preliminary steps to matching.

Additional matching techniques include use of lists – given names, nicknames, etc. – and use of additional attributes such as birthdate when available. Advanced matching techniques include lexical and semantic algorithms.

IDENTITY MATCHING AND RESOLUTION

Identity matching involves recognition of individuals (individual customers, suppliers, accounts, employees, etc.) to support positive identification. Recognition of common identity often uses complex logic involving several data elements and algorithms for semantic similarities and match probability.

Identity resolution determines what actions to take when multiple records are matched and determined to represent a single individual. Resolution is more complex than simply choosing “winner” and “loser” records. It is often necessary to consolidate data by combining columns from multiple records to create a single view of the individual.

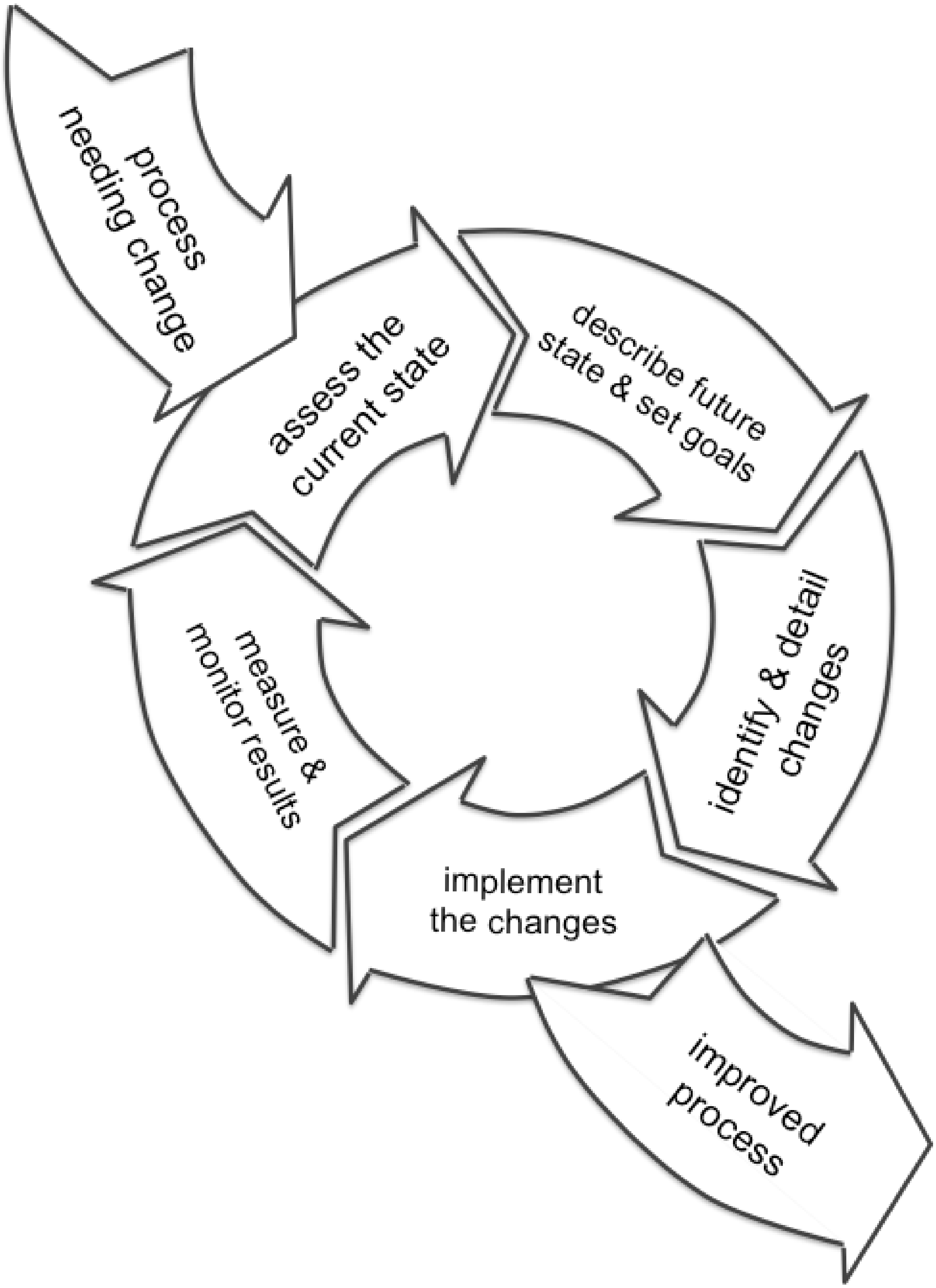
Transforming Data With Intelligence™

Module 5

Preventing Data Quality Defects

Process Improvement

Process Improvement Principles



Process Improvement

Process Improvement Principles

PROCESS IMPROVEMENT DEFINED

Process improvement is the work of preventing occurrence of future defects. In data quality, as with any other product, causes of defects fall into two broad categories – defective materials and process deficiencies. Process improvement focuses on correcting process deficiencies to eliminate causes of defects.

PROCESS IMPROVEMENT CYCLES

Process improvement begins with recognition of a process needing to change, and ends with implementation of an improved process. Between the beginning and the end is a cyclic process of:

- Assess the current state – know where you are objectively
- Describe the future state and set goals – know where you want to go and make it measurable
- Identify and detail changes – build an action plan
- Implement the changes – execute the action plan
- Measure and monitor results – check progress against goals

And repeat the cycle until the process is optimized.