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 Data Quality Management

Techniques for Data Profiling, Assessment, and Improvement

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

info@tdwi.org

Publication Date: February 2012

TABLE OF CONTENTS

COURSE OBJECTIVES

To learn:

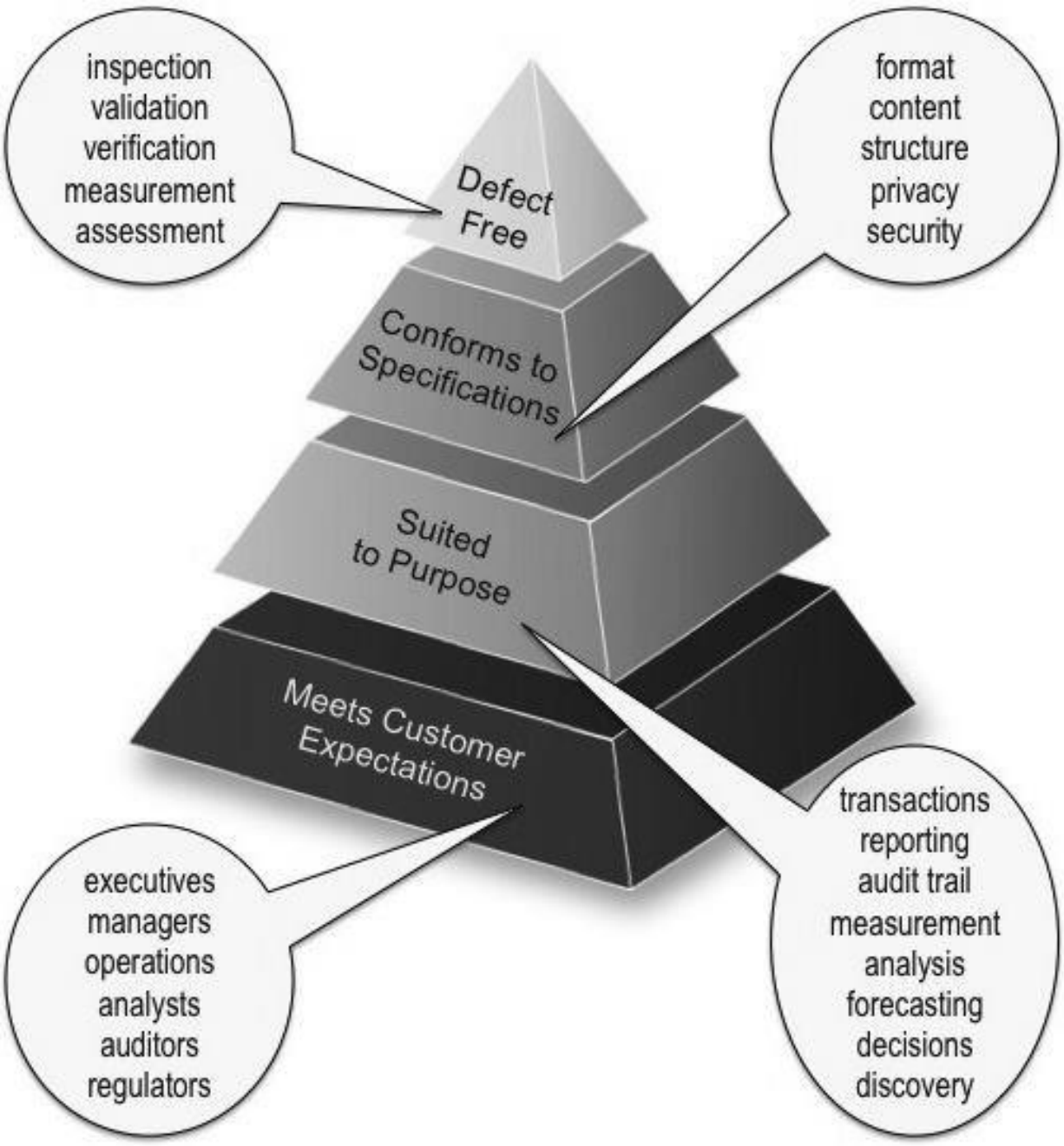
- ✓ ***Techniques for column, table, and cross-table data profiling***
- ✓ ***How to analyze data profiles and find the stories within them***
- ✓ ***Subjective and objective methods to assess and measure data quality***
- ✓ ***How to apply OLAP and performance scorecards for data quality management***
- ✓ ***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***

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

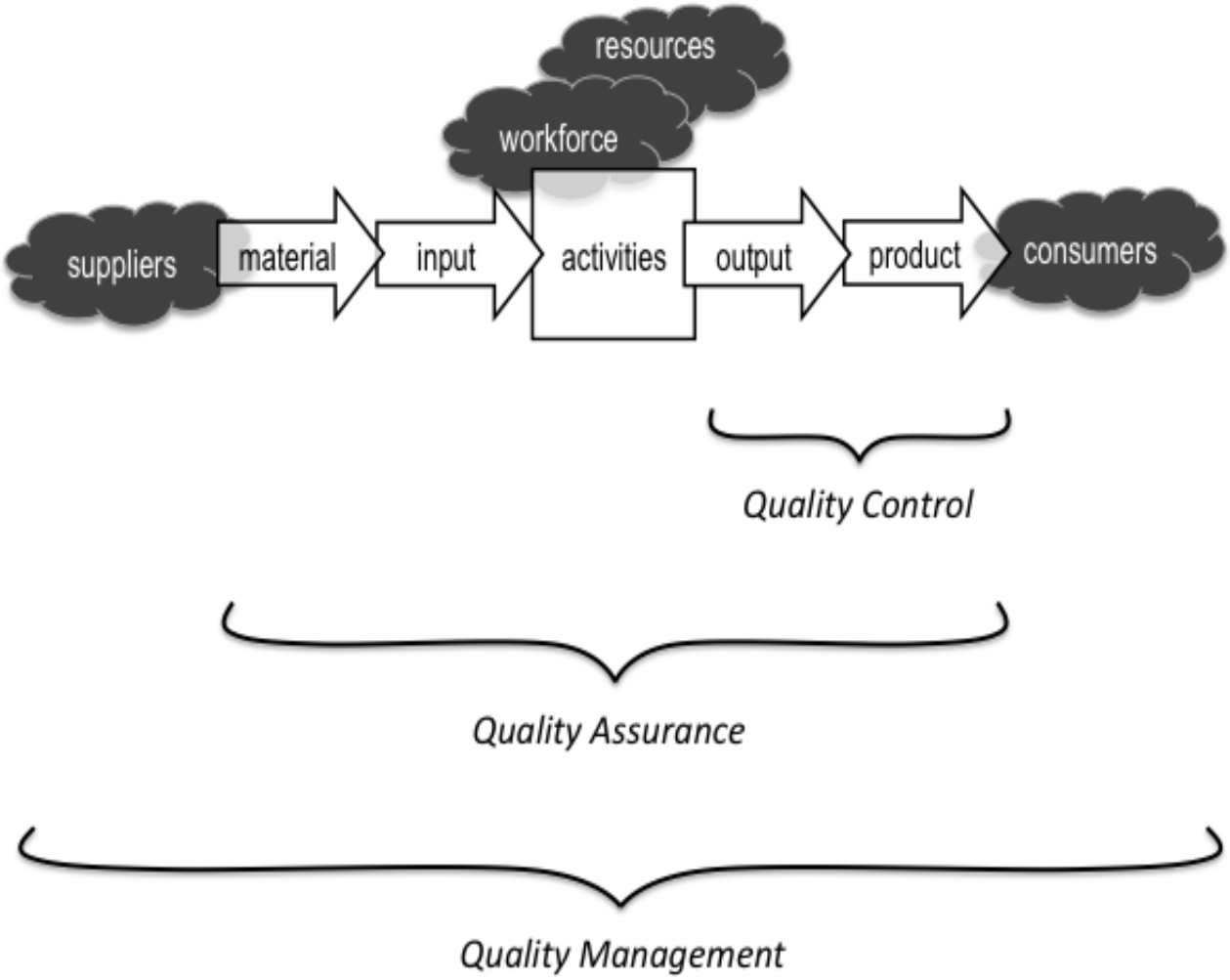
Suitability to purpose must consider all purposes for which data is used, ranging from business transactions and operational reporting to business intelligence and analytics. Expect the quality criteria to vary widely among the different uses. Variations in quality criteria increase the level of difficulty 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:

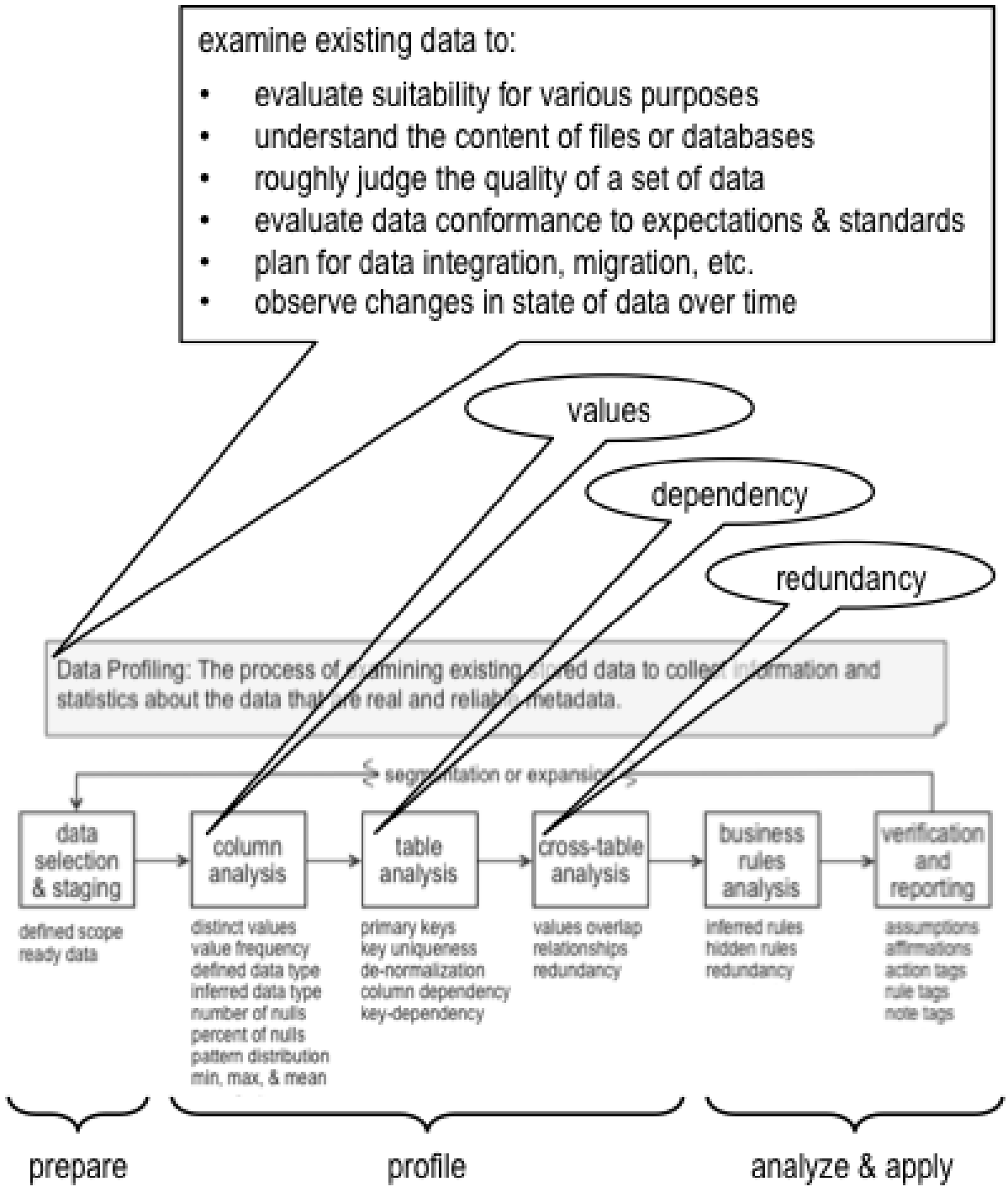
- Quality control (QC) is the narrowest view of QM, and is based on checking the product for defects before it is released.
- Quality assurance (QA) broadens the view by looking "up the line" to check quality at the activities and materials stages of production. QA includes QC and more.
- The end-to-end 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.

Module 2

Profiling Data

Data Profiling Concepts

Purpose and Processes



Data Profiling Concepts

Purpose and Processes

WHY PROFILE?

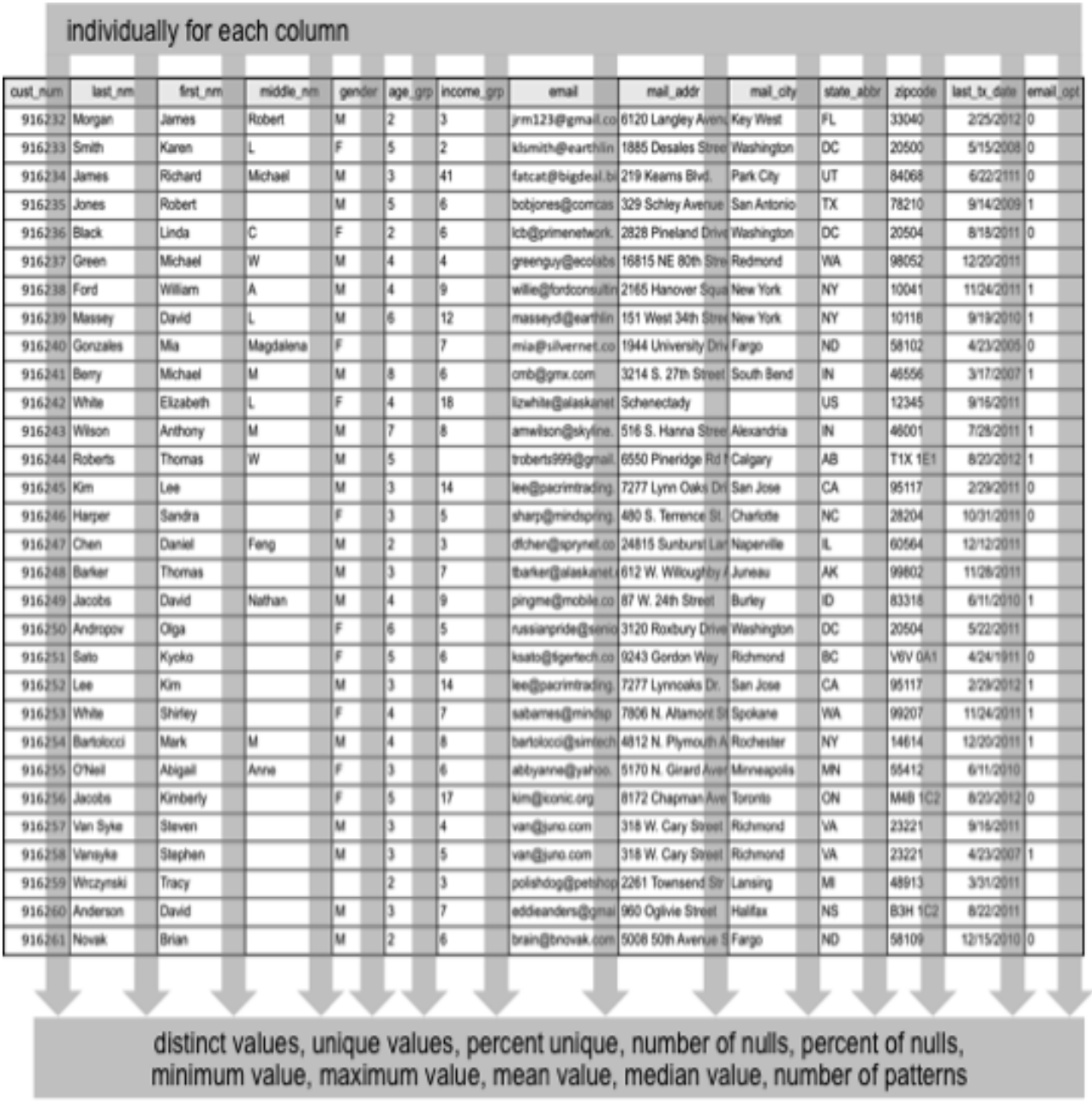
Data profiling is the work of understanding the data by looking at the data. While looking at the data may seem an obvious necessity to some, it is often overlooked. The tendency to review data models, descriptions, definitions, and program code causes many to overlook the obvious. And those who do look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

Column Profiling

Extracting Values Metadata



cust_num	last_nm	first_nm	middle_nm	gender	age_grp	income_grp	email	mail_addr	mail_city	state_abbr	zipcode	last_tx_date	email_opt
916232	Morgan	James	Robert	M	2	3	jrm123@gmail.co	6120 Langley Aven	Key West	FL	33040	2/25/2012	0
916233	Smith	Karen	L	F	5	2	klsmith@earthlin	1885 Desales Stree	Washington	DC	20500	5/15/2008	0
916234	James	Richard	Michael	M	3	41	fatcat@bigdeal.bi	219 Kearns Blvd.	Park City	UT	84068	6/22/2111	0
916235	Jones	Robert		M	5	6	bobjones@comcas	329 Schley Avenue	San Antonio	TX	78210	9/14/2009	1
916236	Black	Linda	C	F	2	6	lcb@primenetwork.	2828 Pineland Drive	Washington	DC	20504	8/18/2011	0
916237	Green	Michael	W	M	4	4	greenguy@ecolabs	16815 NE 80th Stre	Redmond	WA	98052	12/20/2011	
916238	Ford	William	A	M	4	9	willie@fordconsultin	2165 Hanover Squa	New York	NY	10041	11/24/2011	1
916239	Massey	David	L	M	6	12	masseyd@earthlin	151 West 34th Stree	New York	NY	10118	9/19/2010	1
916240	Gonzales	Mia	Magdalena	F		7	mia@silvernet.co	1944 University Driv	Fargo	ND	58102	4/23/2005	0
916241	Berry	Michael	M	M	8	6	cmb@gmx.com	3214 S. 27th Street	South Bend	IN	46556	3/17/2007	1
916242	White	Elizabeth	L	F	4	18	lizwhite@alaskanet	Schenectady		US	12345	9/15/2011	
916243	Wilson	Anthony	M	M	7	8	amwilson@skyline.	516 S. Hanna Stree	Alexandria	IN	46001	7/28/2011	1
916244	Roberts	Thomas	W	M	5		troberts999@gmail.	6550 Pineridge Rd #	Calgary	AB	T1X 1E1	8/22/2012	1
916245	Kim	Lee		M	3	14	lee@pacrimtrading.	7277 Lynn Oaks Dri	San Jose	CA	95117	2/29/2011	0
916246	Harper	Sandra		F	3	5	sharp@mindspring.	480 S. Terrence St.	Charlotte	NC	28204	10/31/2011	0
916247	Chen	Daniel	Feng	M	2	3	dfchen@sprynet.co	24815 Sunburst Lan	Naperville	IL	60564	12/12/2011	
916248	Barker	Thomas		M	3	7	tbarker@alaskanet.c	612 W. Willoughby A	Juneau	AK	99802	11/28/2011	
916249	Jacobs	David	Nathan	M	4	9	pingme@mobile.co	87 W. 24th Street	Burley	ID	83318	6/11/2010	1
916250	Andrrpov	Olga		F	6	5	russianpride@senio	3120 Roxbury Drive	Washington	DC	20504	5/22/2011	
916251	Sato	Kyoko		F	5	6	ksato@tigertech.co	9243 Gordon Way	Richmond	BC	V6V 0A1	4/24/1911	0
916252	Lee	Kim		M	3	14	lee@pacrimtrading.	7277 Lynnoaks Dr.	San Jose	CA	95117	2/29/2012	1
916253	White	Shirley		F	4	7	sabarnes@mindsp	7806 N. Altamont St	Spokane	WA	99207	11/24/2011	1
916254	Bartolocci	Mark	M	M	4	8	bartolocci@simtech	4812 N. Plymouth A	Rochester	NY	14614	12/20/2011	1
916255	O'Neil	Abigail	Anne	F	3	6	abbyanne@yahoo.	5170 N. Girard Ave	Minneapolis	MN	55412	6/11/2010	
916256	Jacobs	Kimberly		F	5	17	kim@iconic.org	8172 Chapman Ave	Toronto	ON	M4B 1C2	8/20/2012	0
916257	Van Syke	Steven		M	3	4	van@juno.com	318 W. Cary Street	Richmond	VA	23221	9/15/2011	
916258	Vansyka	Stephen		M	3	5	van@juno.com	318 W. Cary Street	Richmond	VA	23221	4/23/2007	1
916259	Wrczynski	Tracy			2	3	polishdog@petshop	2261 Townsend Str	Lansing	MI	48913	3/31/2011	
916260	Anderson	David		M	3	7	eddieanders@gmai	960 Oglivie Street	Halifax	NS	B3H 1C2	8/22/2011	
916261	Novak	Brian		M	2	6	brain@bnovak.com	5008 50th Avenue S	Fargo	ND	58109	12/15/2010	0

Column Profiling

Extracting Values Metadata

FREQUENCY OF VALUES

Frequency of values, or frequency distribution, is a statistical concept that is central to data profiling. Frequency distribution is the tabulation of the number of occurrences of each unique value in a data set; in profiling it tabulates unique values in a column of data. Frequency distribution is the foundation from which other column metatdata is determined.

BUILDING ON THE FOUNDATION

Statistics are a necessary foundation from which much more metadata can be derived. Profiling a column of data produces obvious statistics such as min, max, mean, frequency, and a list of distinct values. That metadata, however, is only the beginning. It can be applied to derive and infer much more:

- For a single column it is possible to infer a data type from the set of unique values, and to compare the inferred data type with the data type that is defined for the column. You might, for example, profile a column defined as VCHAR and find that all of the values in that column appear to be DATE data.

- Additional inference is possible when profiling multiple columns of a table. A column has unique values, for example, when frequency distribution is flat and the number of distinct values is equal to the count of all values. Probability is high that the column is a primary key for the table.

- Patterns of correspondence among columns in a single table lead to inference of column dependency, discovery of multi-valued columns, evidence of repeating groups (embedded tables), and other forms of de-normalization.

- Profiling of columns across multiple tables exposes overlapping domains of values which are indicators of table-to-table relationships and of data redundancy.

Table Profiling

Examining Dependencies

official abbreviation for designated postal area

type of designated postal area

minimum zip code value for US postal designations

maximum zip code value for US postal designations

first letter of postal code for Canada provinces

name or description of designated postal area

abbr	type	zipcode_low	zipcode_high	ca_prefix	name
AA	US Military	34001	34099		ARMED FORCES - AMERICA
AB	Canada Province			T	ALBERTA
AE	US Military	09003	09898		ARMED FORCES - AFRICA, CANADA, EUROPE, MIDDLE EAST)
AK	US State	99501	99950		ALASKA
AL	US State	35004	36925		ALABAMA
AP	US Military	96200	96299		ARMED FORCES PACIFIC
AR	US State	71601	72959		ARKANSAS
AS	US Territory	96700	96799		AMERICAN SAMOA
AZ	US State	85001	86556		ARIZONA
BC	Canada Province			V	BRITISH COLUMBIA
CA	US State	90001	96962		CALIFORNIA
CO	US State	80001	81658		COLORADO
CT	US State	06001	06928		CONNECTICUT
DC	US District	20001	20599		DISTRICT OF COLUMBIA
DE	US State	19701	19980		DELAWARE
FL	US State	32004	34997		FLORIDA
FM	US Territory	96900	96999		FEDERATED STATES OF MICRONESIA
GA	US State	30002	39901		GEORGIA
WI	US State	53001	54990		WISCONSIN
WV	US State	24701	26886		WEST VIRGINIA
WY	US State	82001	83128		WYOMING
YT	Canada Province			Y	YUKON

Table Profiling

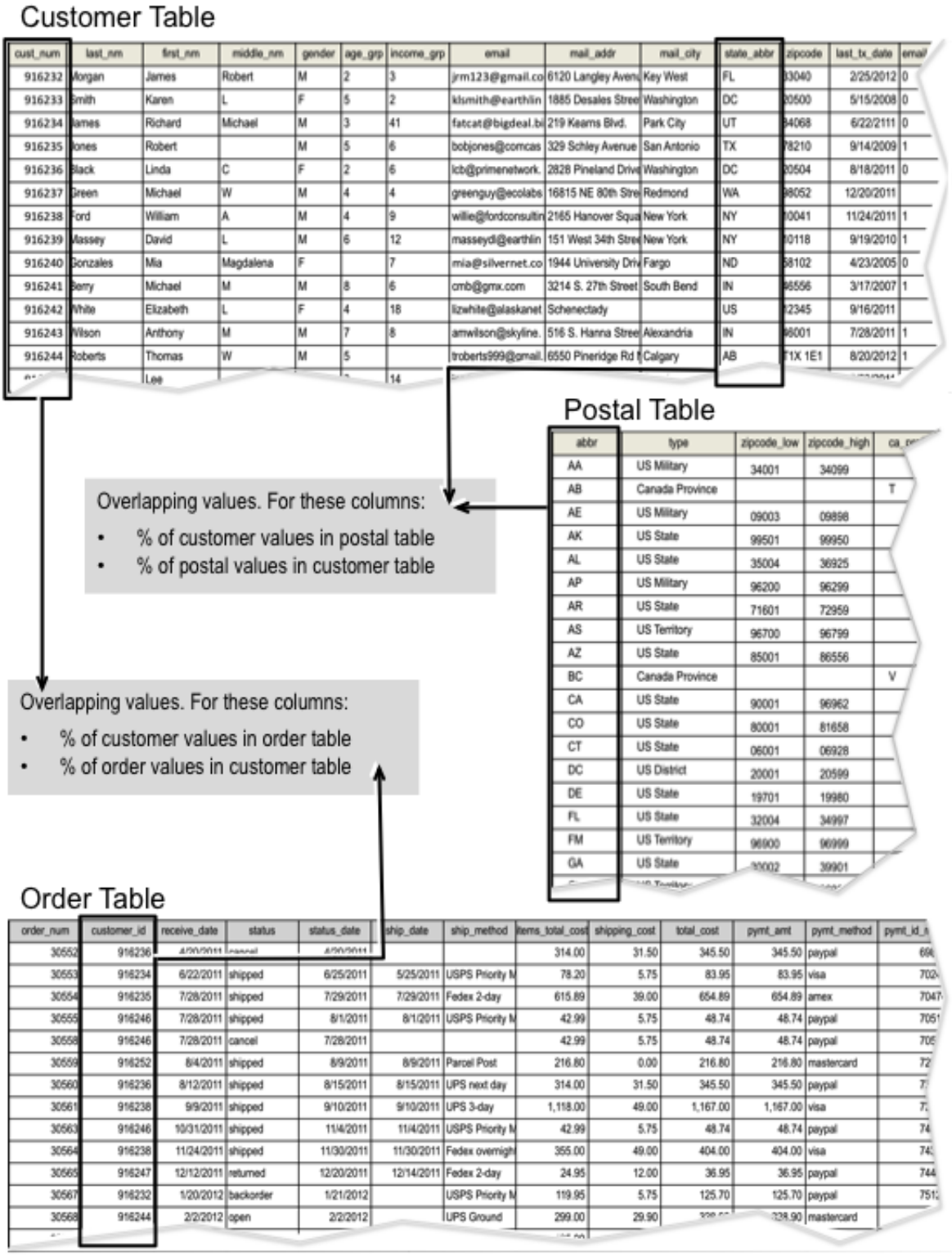
Examining Dependencies

LOOKING AT MULTIPLE COLUMNS

Where column profiling examines one column at a time, table profiling looks for relationships and dependencies among the columns of a table. We'll begin the table profiling examples using the postal table that is illustrated on the facing page.

Cross-Table Profiling

Examining Redundancy and Relationships



Cross-Table Profiling

Examining Redundancy and Relationships

VALUES OVERLAP ACROSS TABLES

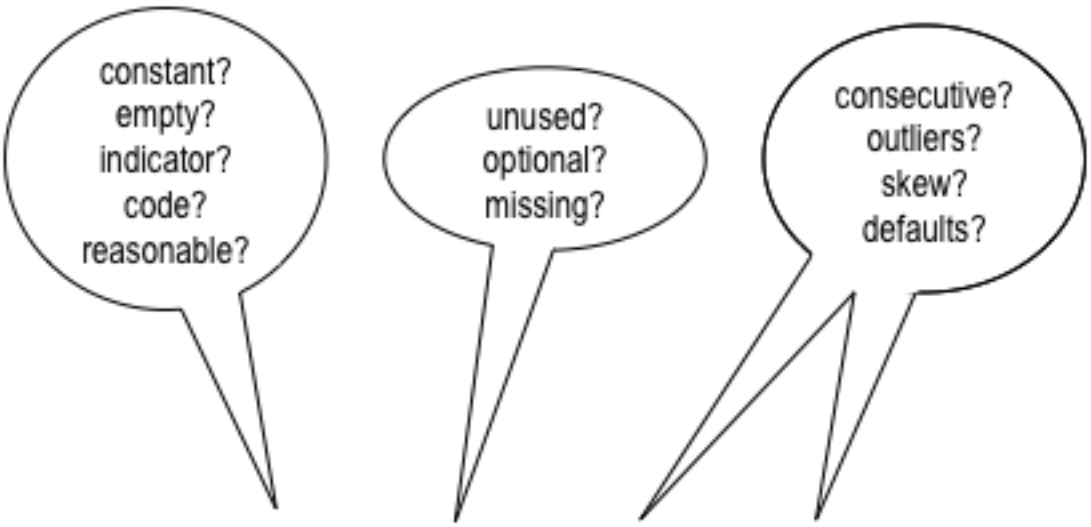
Cross-table profiling examines relationships between columns in different tables. Through cross-table profiling you may find foreign key relationships, redundancy, inconsistency, synonymous but differently named columns, similarly named columns with circumstantial differences, and more.

HOW IT WORKS

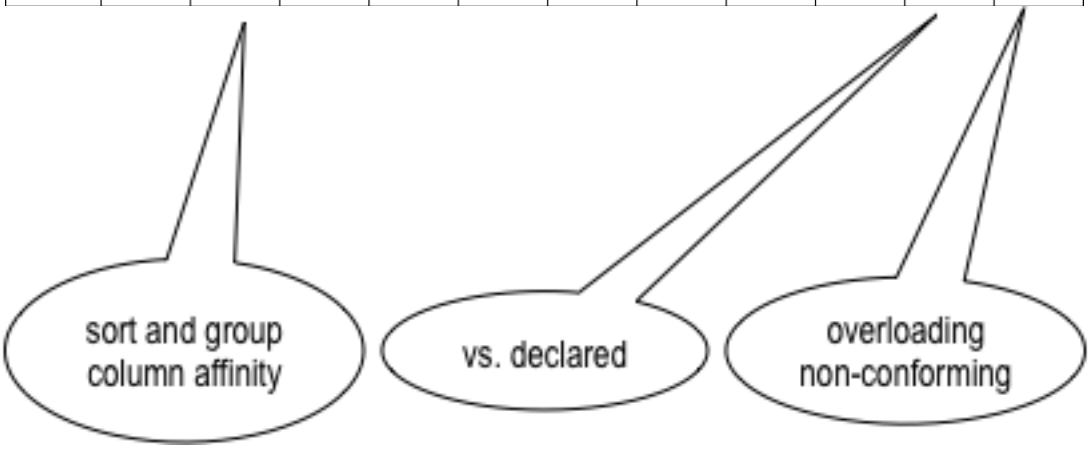
Profiling tools build a distinct values list for each column when column profiling is done. Cross-table profiling compares each distinct values list with every other distinct values list to find those columns where significant overlap of values occurs. Of course, you must determine the overlaps that are interesting and meaningful. Dates or indicators, for example, are likely to have high levels of overlap even when there is little or no redundancy of information.

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

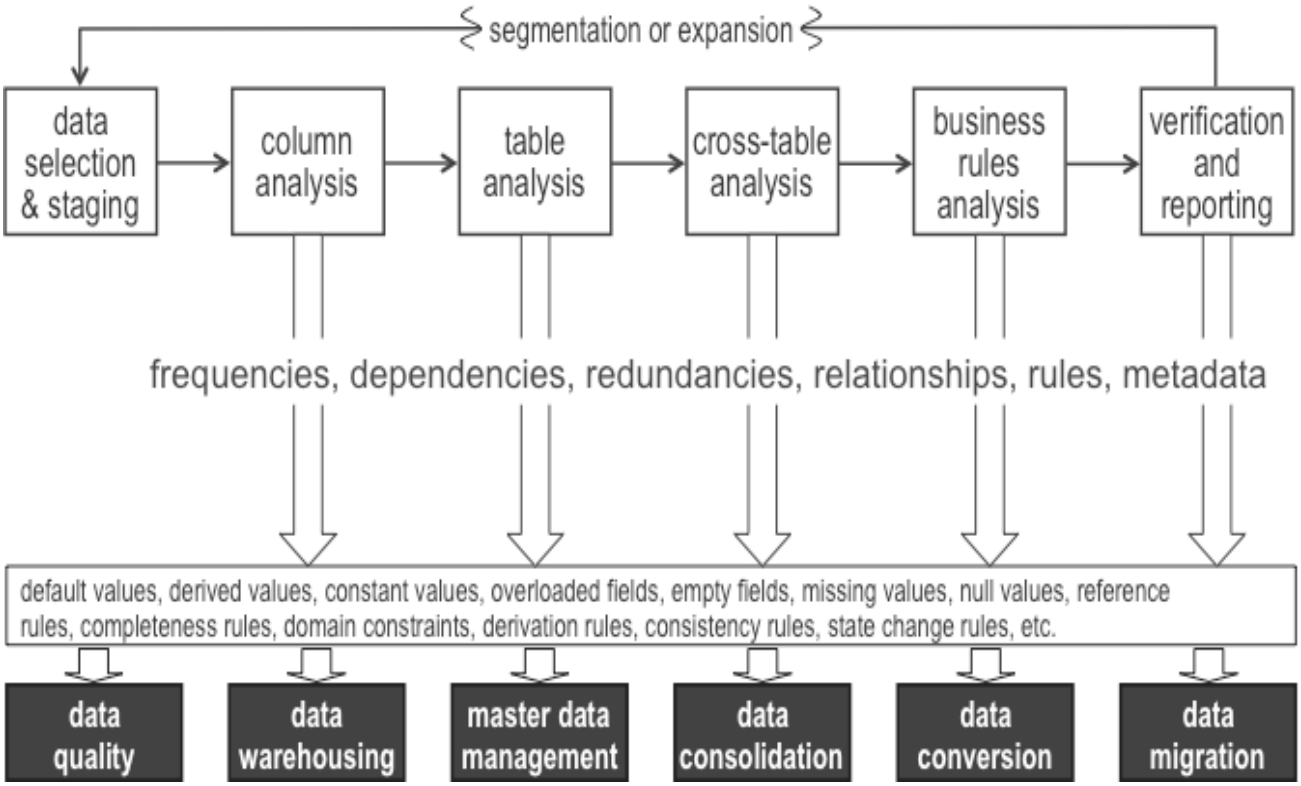
METADATA MATCHING

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)

Data Profiling in Practice

Profiling and Projects



Data Profiling in Practice

Profiling and Projects

THE NEED TO PROFILE DATA

Rarely do we undertake data management projects where data profiling doesn't contribute value. Profiling has direct impact on data quality through data quality projects. It has less direct but very real impact with many other kinds of projects including:

- Data warehousing
- Master data management
- Data consolidation
- Data conversion
- Data migration

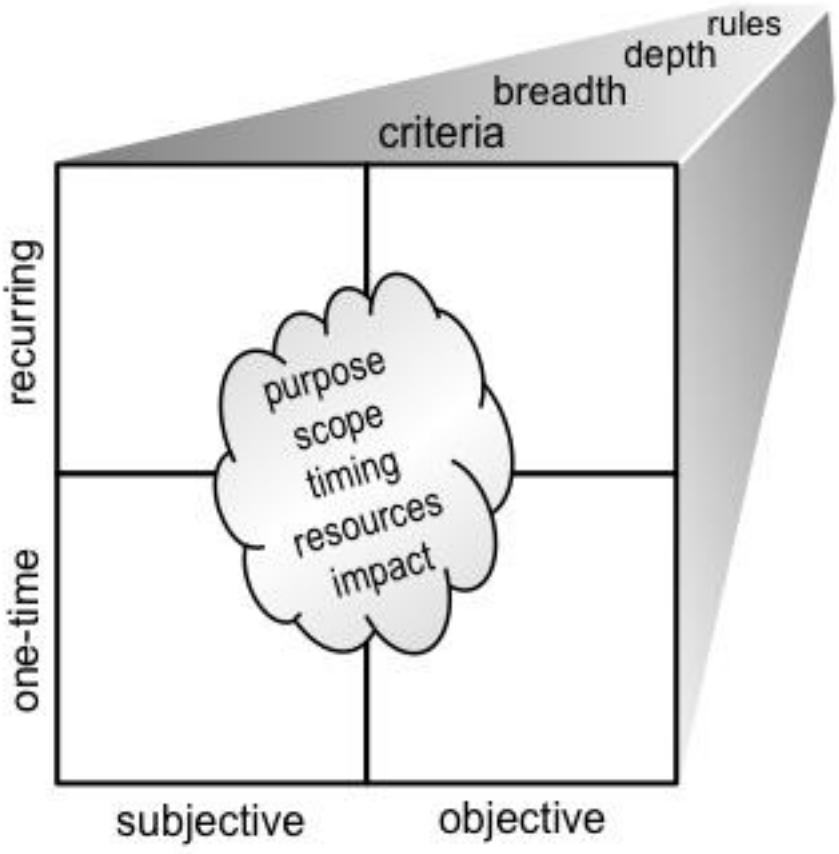
Many data quality practitioners believe that consolidation, conversion, and migration are key areas where we introduce data quality defects. If that is true, it is because we perform the work without fully understanding the data. Data profiling can make a real difference to these projects.

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.

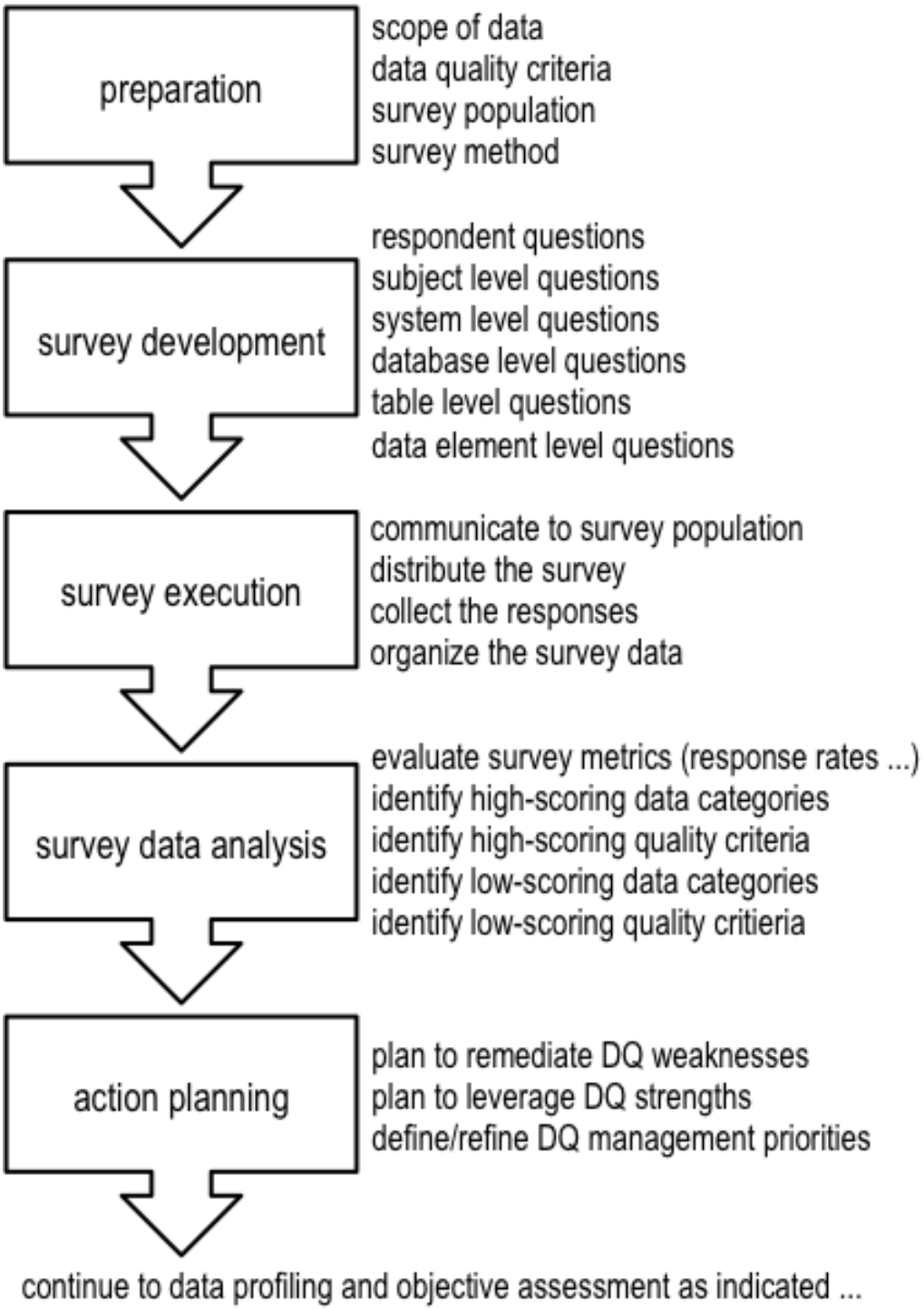
Choosing the type (or types) of assessment – one-time or recurring, subjective or objective – is guided by several factors including:

- 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.

Subjective Assessment

Subjective Assessment Process



Subjective Assessment

Subjective Assessment Process

SURVEY AND ANALYSIS

Subjective assessment is a survey and analysis process that consists of five steps as illustrated on the facing page.

PREPARATION

Preparation is an especially important step in any survey-driven process. Good surveys are difficult and bad surveys are easy. The preparation steps help to ensure that you get good and useful data by focusing on

- Scope of data – broad enough to collect data that is informative yet narrow enough that it is significant to the survey population. Also consider the effect of scope on the size of the survey.
- Data quality criteria – five or six key criteria may yield better results than a survey with 15 or 20 criteria. Choose a relatively small set of the quality criteria that are most important right now.
- Survey population is ideally a cross section of people with interest in quality for the scope of data – business and technical, management and functional staff, recent hires and long-term employees, etc.
- Survey method fits together with survey population to drive response rates. For a large and geographically distributed population, email and internet surveys work well. For smaller, localized populations a paper or spreadsheet survey can work well.

SURVEY DEVELOPMENT AND EXECUTION

A good survey collects data about the respondents as well as beliefs about data quality. Respondent demographics provide some of the dimensions for analysis of results. Quality-specific questions connect the scope of data with specific quality criteria. A five-point Likert scale for each data and criterion combination is the most basic survey method.

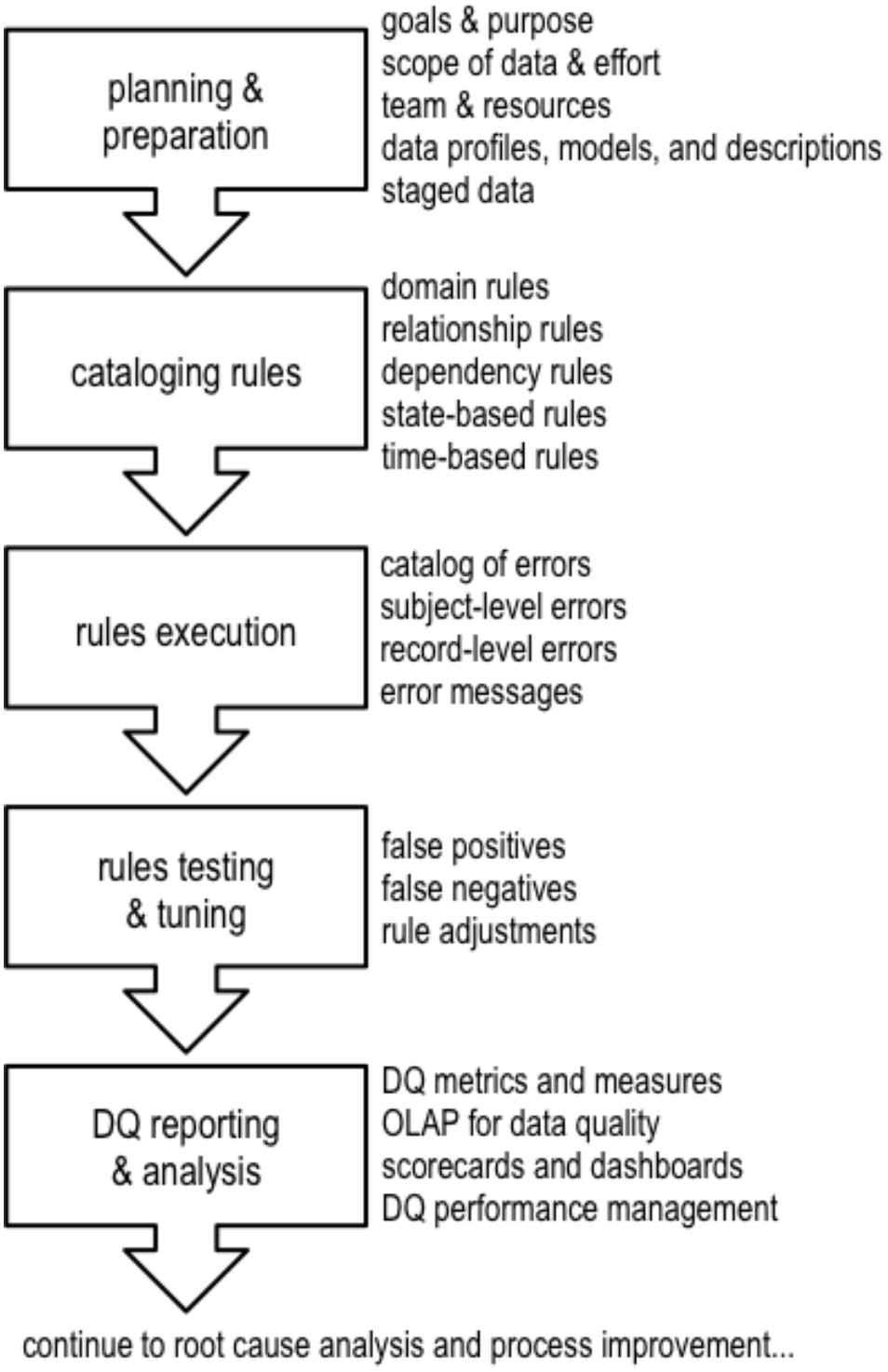
When distributing the survey, communication is important. Let the population know the purpose of the survey, expected time to complete, due date for responses, and what they can expect to see when the data is collected and analyzed. (Feedback is a great motivator to respond.)

ANALYSIS AND ACTION PLAN

Evaluate and analyze the data dimensionally. Look at high scores as well as low scores – good news as well as bad. Also look for divergent responses – managers, for example, giving high rating to an area that functional staff rated poorly. Analysis should identify DQ strengths and weaknesses. Planning sets priorities to remediate weaknesses and to leverage strengths.

Objective Assessment

Objective Assessment Process



Objective Assessment

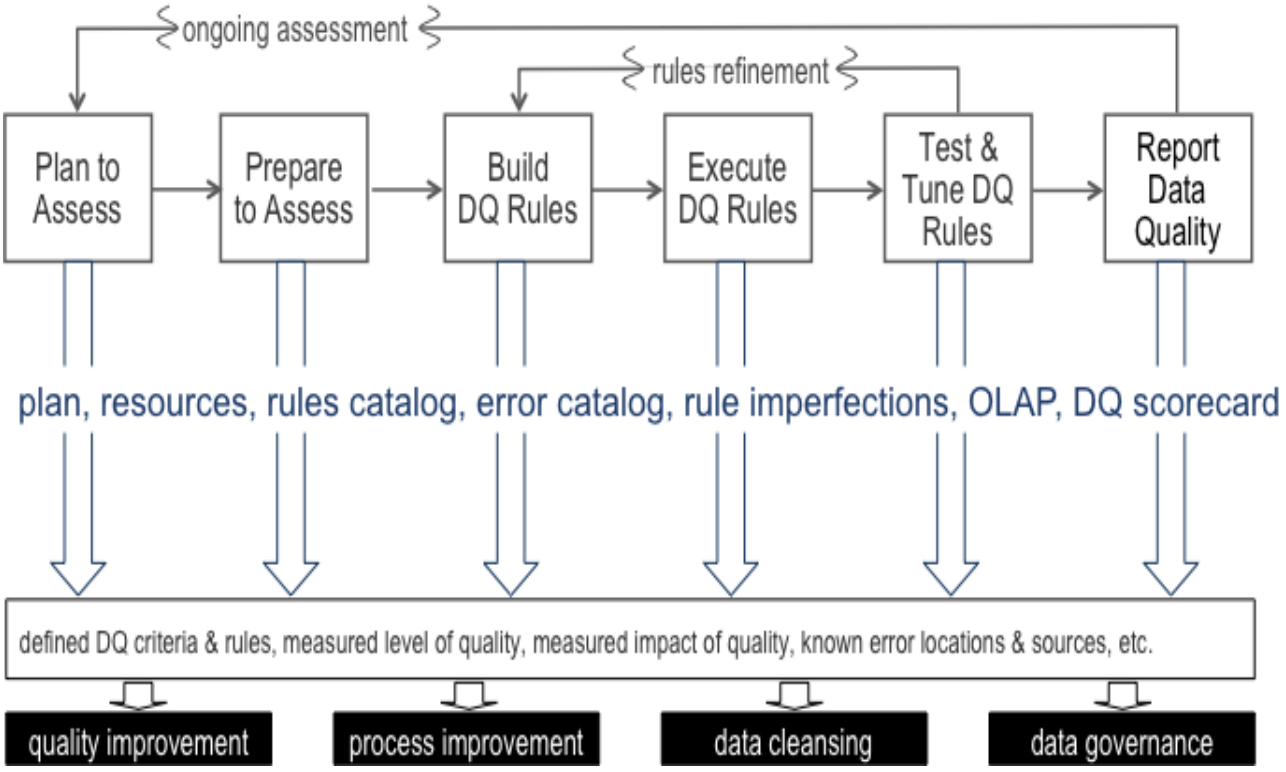
Objective Assessment Process

RULE-BASED DQ ASSESSMENT

Objective data quality assessment is substantially more complex than subjective assessment. Objective assessment encompasses five steps as illustrated on the facing page. Note that three of the five steps are related to data quality rules. Objective assessment is a rule-based process. The work of identifying, recording, executing, and refining data quality rules is much of the effort.

Assessment in Practice

Assessment and Projects



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.

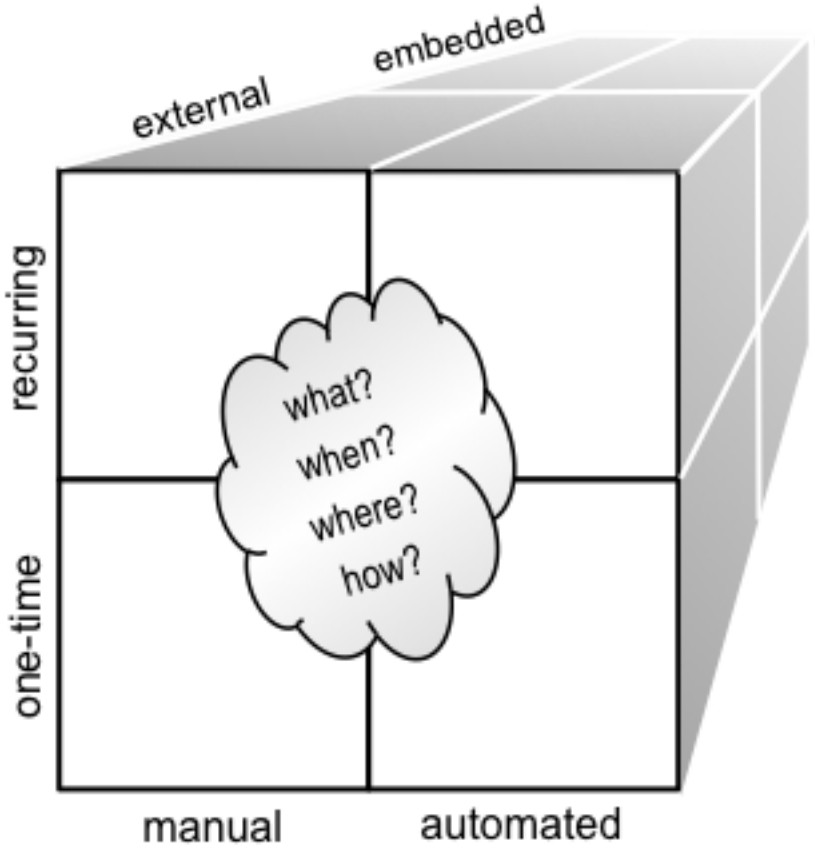
Although not a project but an ongoing program, data governance activities also benefit from data quality assessment. Effective governance requires feedback. For a quality-focused data governance program, assessment produces the feedback that is needed.

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).

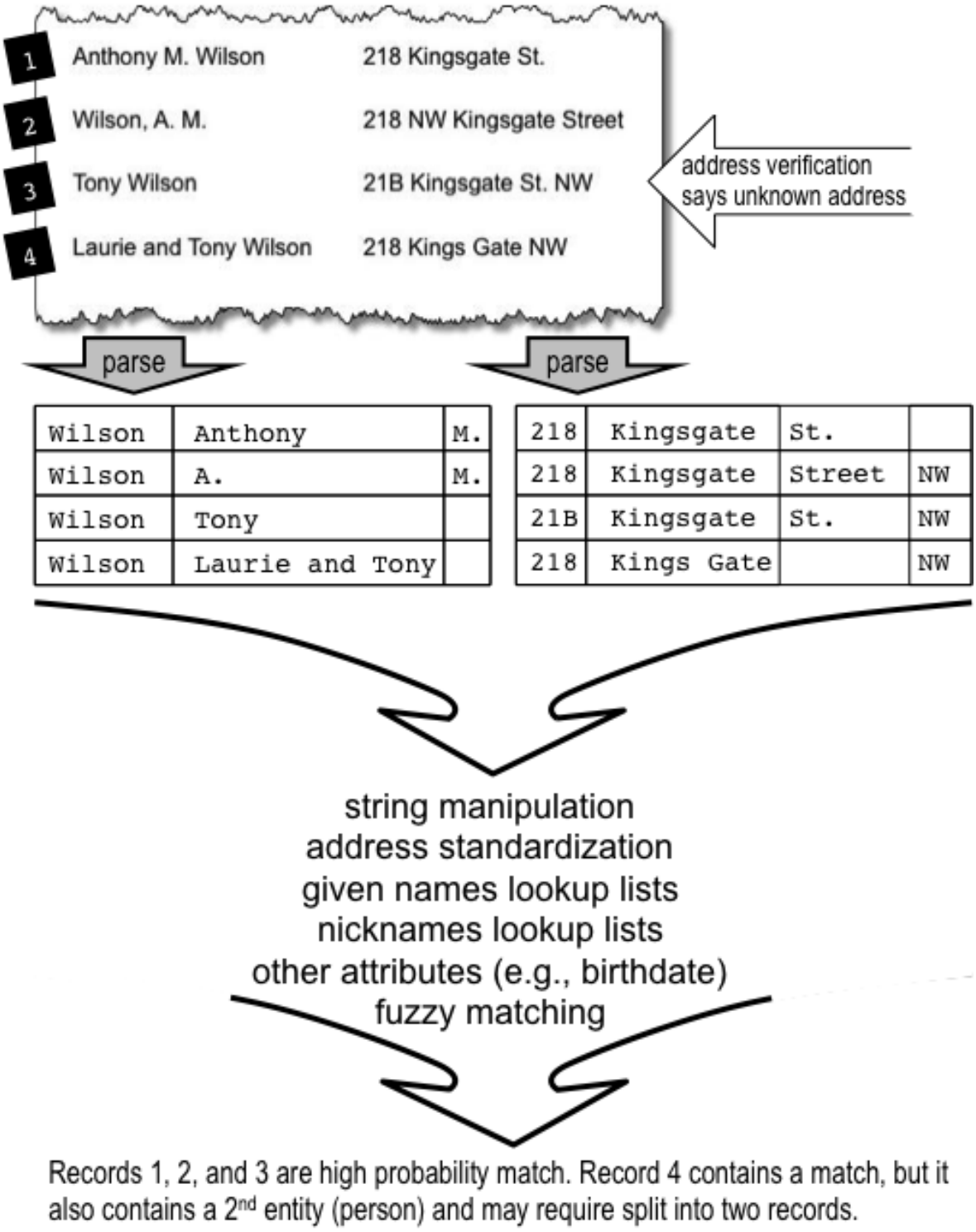
These options combine in some interesting ways – embedded, automated, recurring for example; or external, manual, one-time. A complete data cleansing solution typically uses a mix-and-match approach with several options.

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



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.

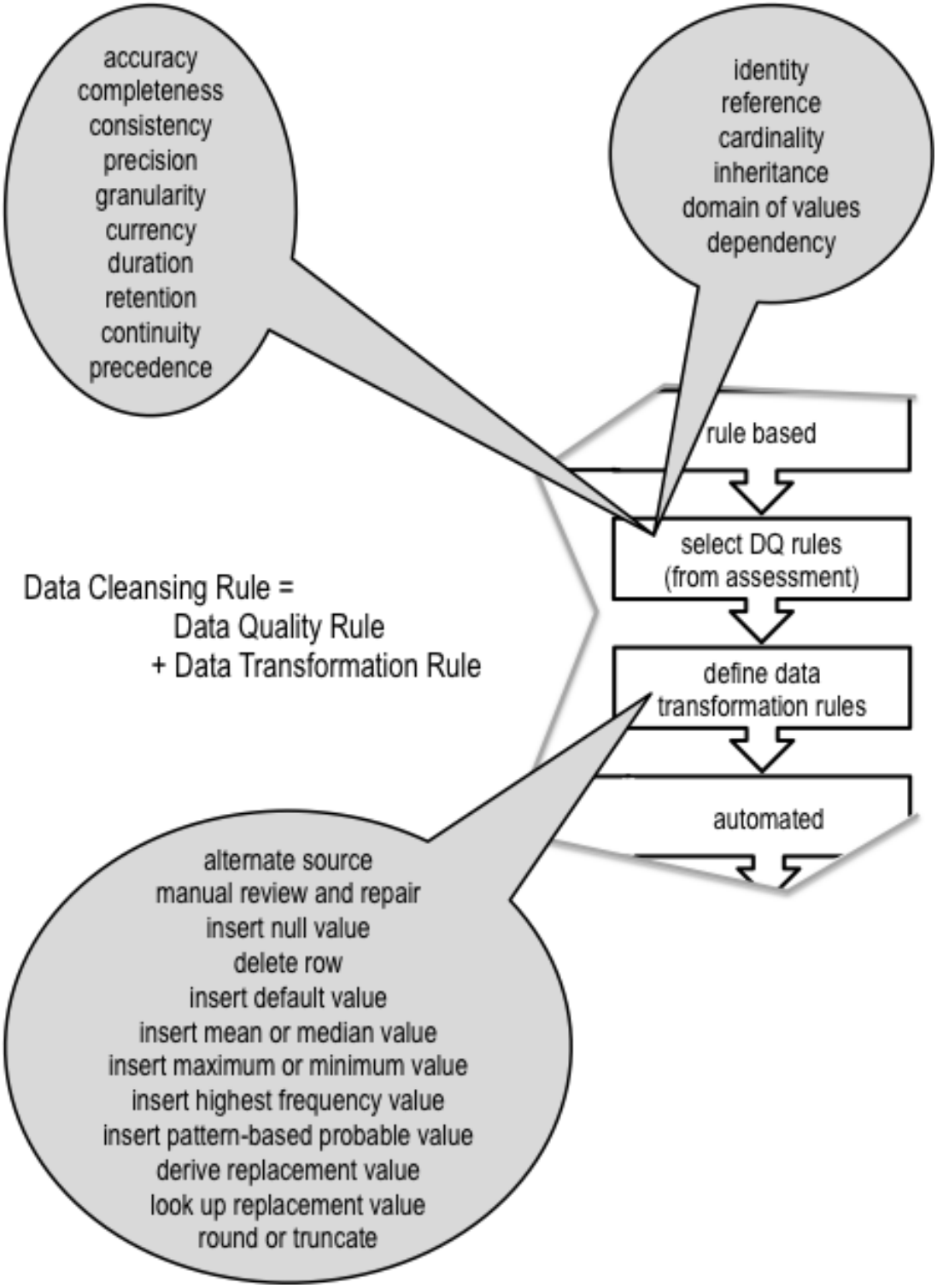
IDENTITY MATCHING AND RESOLUTION

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.

Rule-Based Data Cleansing

Data Cleansing Rules



Rule-Based Data Cleansing

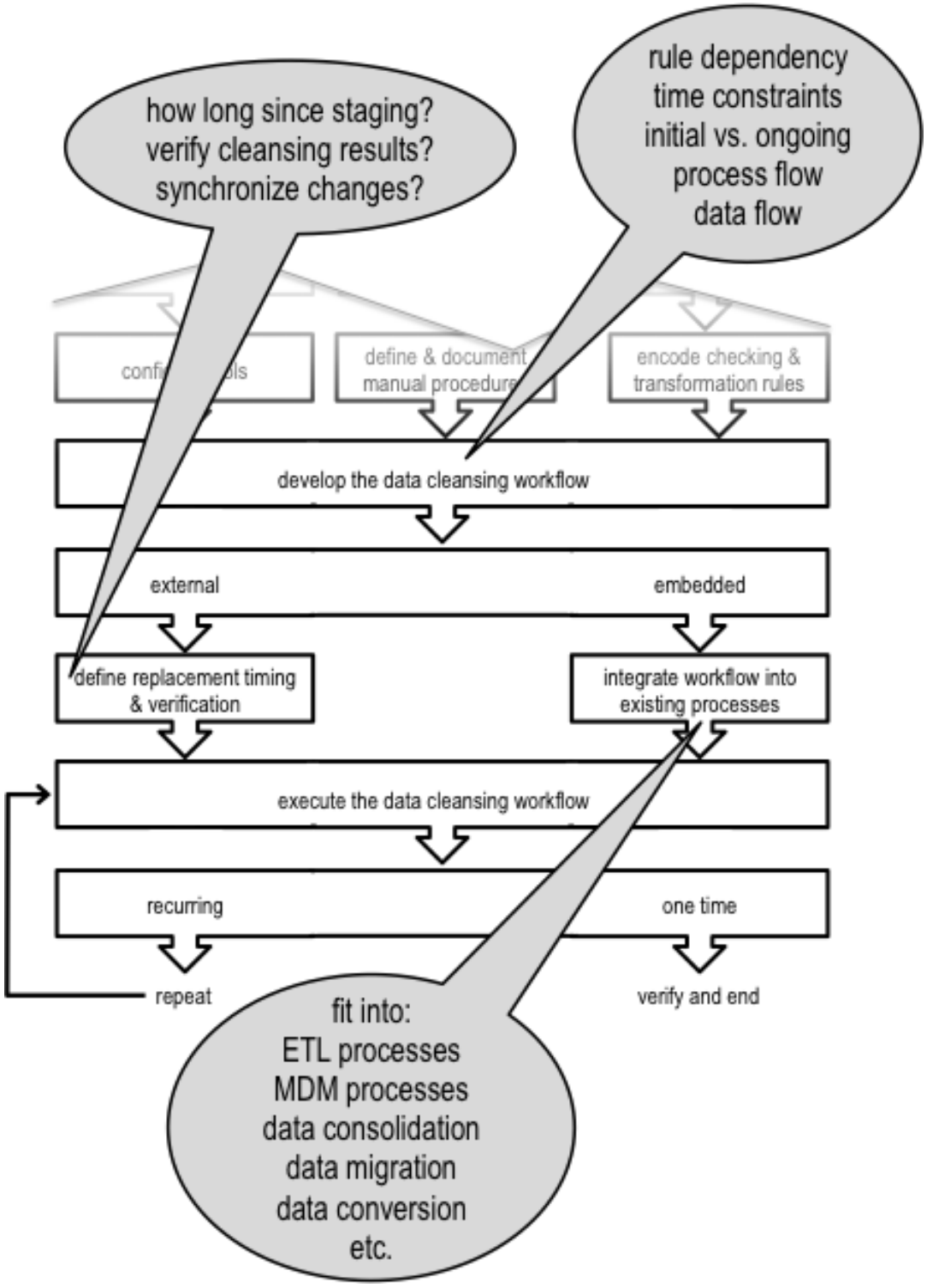
Data Cleansing Rules

BUILDING THE RULES

A data cleansing rule is the combination of one data quality rule with one data transformation rule. Data quality rules are those that we've already discussed expressing integrity and correctness constraints. Data transformation rules describe how data can be changed to improve data quality.

Data Cleansing in Practice

Data Cleansing Workflow



Data Cleansing in Practice

Data Cleansing Workflow

PUTTING THE PIECES TOGETHER

At this point in the process, data cleansing design can seem confusing and a bit overwhelming. You have lots of different pieces:

- One-time cleansing processes
- Recurring cleansing processes
- Procedural techniques
- Rule-based techniques
- Automated processes
- Manual processes
- Embedded processing
- External processing
- Staged data flow
- In place data flow
- ETL data flow

To fit all of the pieces together you need to design a data cleansing workflow with attention to sequence, dependencies, timing, flow, synchronization, and integration into existing processes that simultaneously work with the data.

SYSTEMS DESIGN

You can't do all cleansing at one time, or in one place, or using only one method. Your data cleansing system will mix and match several methods, and processing sequence matters. There is no cookbook or recipe for data cleansing workflow. It is a systems design process much like designing any other system. Data subject matter experts and data quality specialists can contribute, but experienced and skilled systems designers should lead this design activity.

Module 5

Preventing Data Quality Defects

Root Cause Analysis

RCA Overview

Root Cause Analysis (RCA) is a systematic problem solving approach intended to identify root causes of problems or events. RCA is based on the principle that problems are best solved by correcting or eliminating root causes, as opposed to simply addressing symptoms.

- Systematic cause-and-effect analysis
- Address problems – not just symptoms
- More likely that problems will not reoccur
- May require multiple corrective actions
- Often iterative and a component of continuous improvement

From *Root Cause Analysis for Data Quality Management,* © David L. Wells. Reprinted with permission.

Root Cause Analysis

RCA Overview

WHAT IS RCA?

Root cause analysis (RCA) is a systematic problem solving approach intended to identify root causes of problems or events. RCA is based on the principle that problems are best solved by correcting or eliminating root causes, and not simply by fixing symptoms.

There are two key points in this definition:

- RCA is systematic analysis. It has structure, steps, and process.
- The goal is to get beyond symptoms and find real causes.

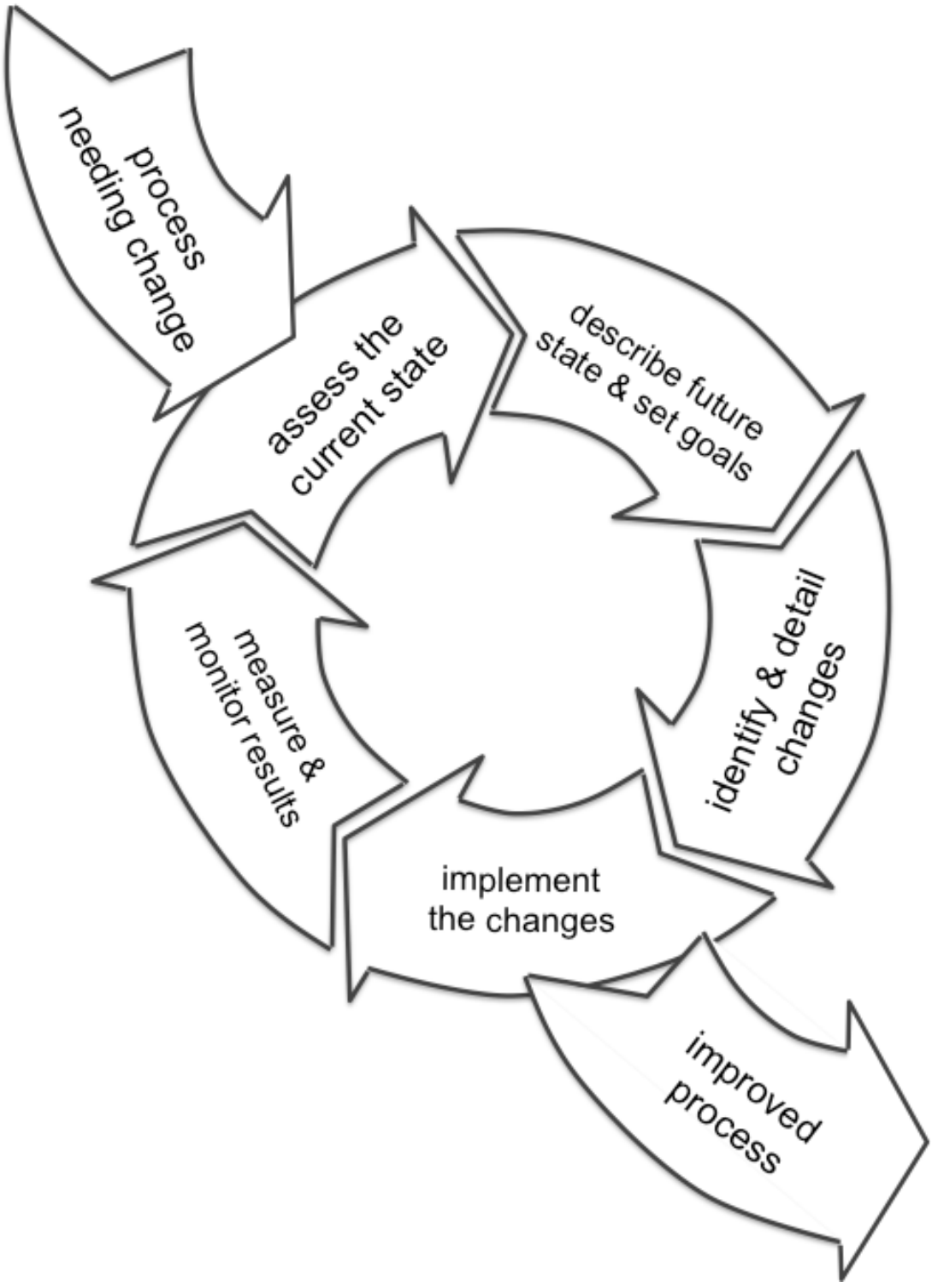
When we address root causes instead of symptoms, we increase the probability that problems will not reoccur.

RCA may require multiple cycles of corrective action. Finding and correcting a root cause may result in the problem shifting to another location in the system or lead to unexpected side effects or consequences. In these instances we need to perform further analysis and repeat the process.

Root cause analysis is often iterative, often cyclical, and an integral part of continuous improvement processes and activities.

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.

