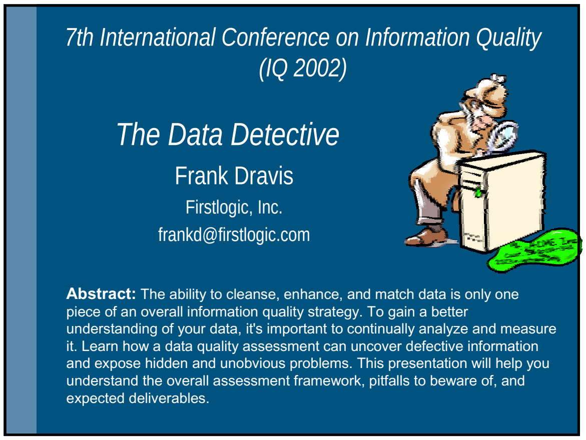
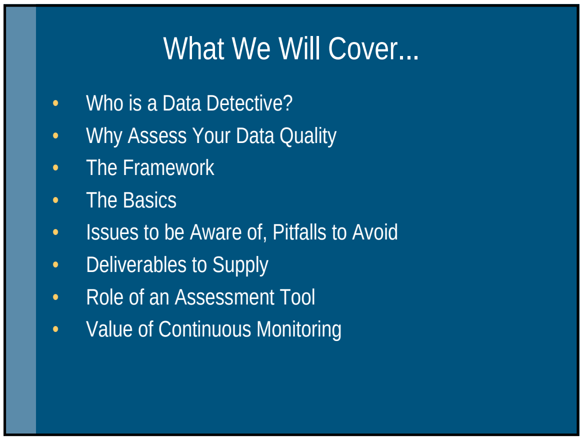


*7th International Conference on Information Quality (IQ 2002)*

## *The Data Detective*

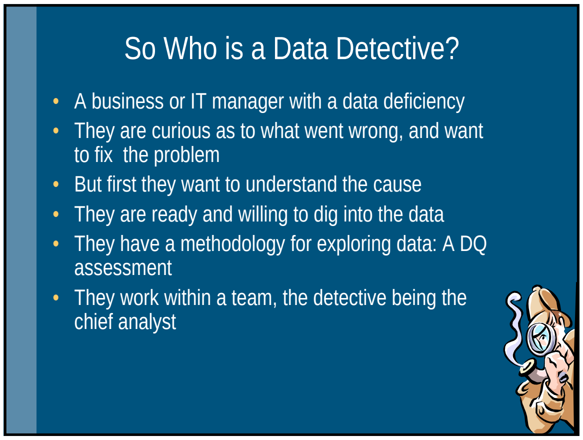
Frank Dravis
Firstlogic, Inc.
frankd@firstlogic.com

**Abstract:** The ability to cleanse, enhance, and match data is only one piece of an overall information quality strategy. To gain a better understanding of your data, it's important to continually analyze and measure it. Learn how a data quality assessment can uncover defective information and expose hidden and unobvious problems. This presentation will help you understand the overall assessment framework, pitfalls to beware of, and expected deliverables.



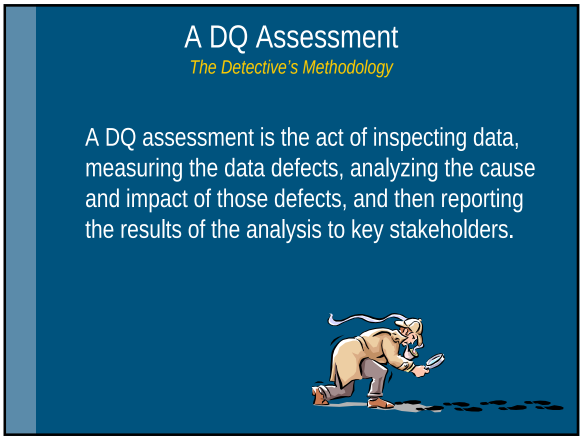
## What We Will Cover…

- Who is a Data Detective?
- Why Assess Your Data Quality
- The Framework
- The Basics
- Issues to be Aware of, Pitfalls to Avoid
- Deliverables to Supply
- Role of an Assessment Tool
- Value of Continuous Monitoring



## So Who is a Data Detective?

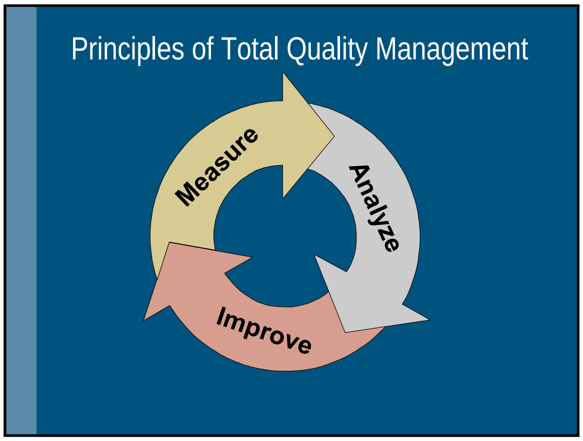
- A business or IT manager with a data deficiency
- They are curious as to what went wrong, and want to fix  the problem
- But first they want to understand the cause
- They are ready and willing to dig into the data
- They have a methodology for exploring data: A DQ assessment
- They work within a team, the detective being the chief analyst



## A DQ Assessment

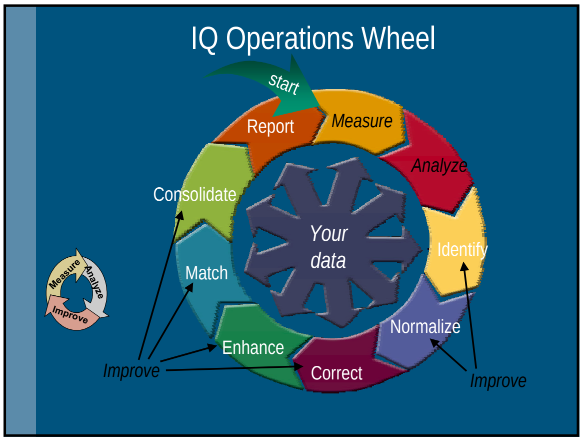
*The Detective's Methodology*

A DQ assessment is the act of inspecting data, measuring the data defects, analyzing the cause and impact of those defects, and then reporting the results of the analysis to key stakeholders.



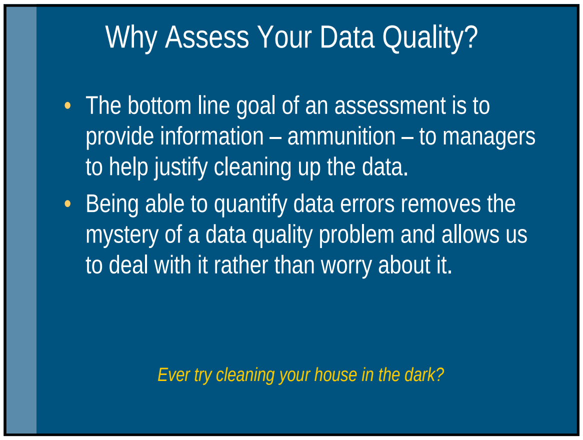
## Principles of Total Quality Management

Measure
Analyze
Improve



## IQ Operations Wheel

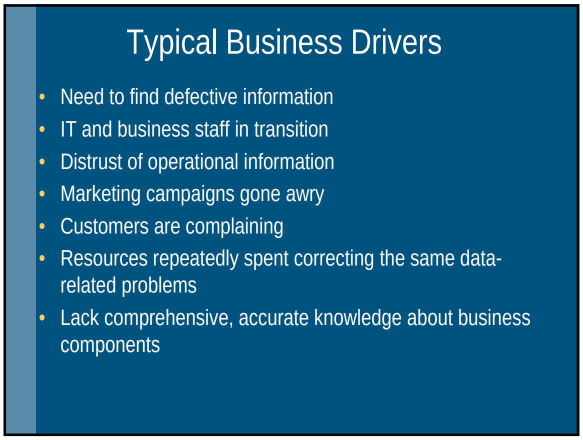
start
Report
Measure
Analyze
Identify
Normalize
Correct
Enhance
Match
Consolidate
*Your data*
Measure
Analyze
Improve
Improve
Improve



## Why Assess Your Data Quality?

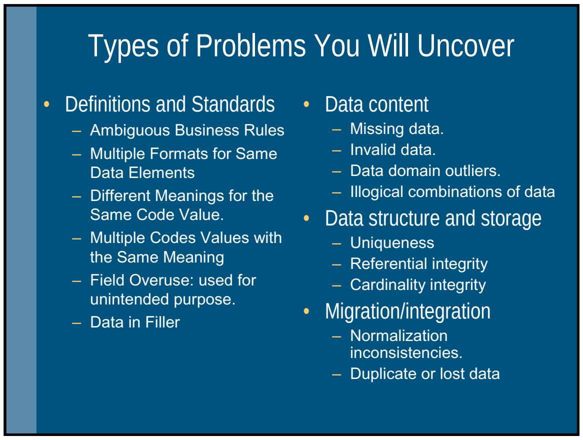
- The bottom line goal of an assessment is to provide information – ammunition – to managers to help justify cleaning up the data.
- Being able to quantify data errors removes the mystery of a data quality problem and allows us to deal with it rather than worry about it.

*Ever try cleaning your house in the dark?*



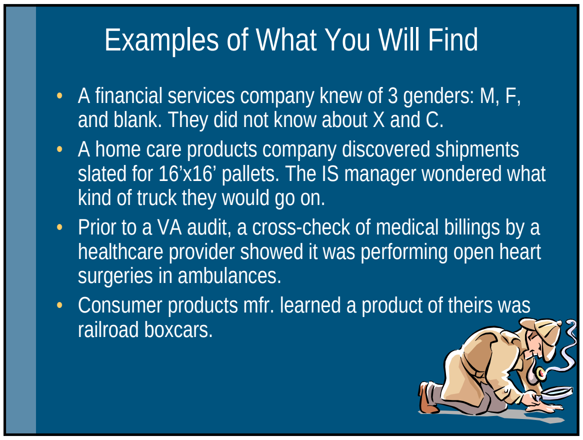
## Typical Business Drivers

- Need to find defective information
- IT and business staff in transition
- Distrust of operational information
- Marketing campaigns gone awry
- Customers are complaining
- Resources repeatedly spent correcting the same data-related problems
- Lack comprehensive, accurate knowledge about business components



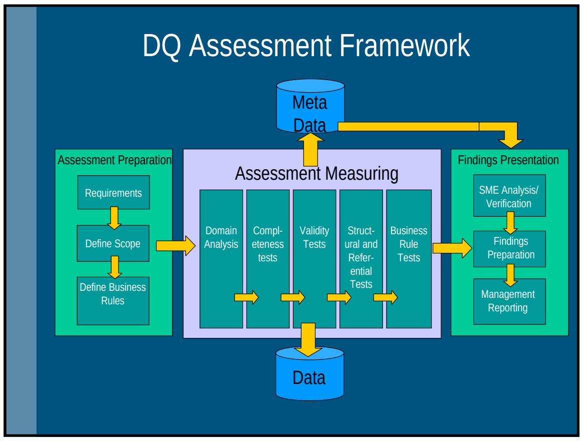
## Types of Problems You Will Uncover

- Definitions and Standards
  - Ambiguous Business Rules
  - Multiple Formats for Same Data Elements
  - Different Meanings for the Same Code Value.
  - Multiple Codes Values with the Same Meaning
  - Field Overuse: used for unintended purpose.
  - Data in Filler
- Data content
  - Missing data.
  - Invalid data.
  - Data domain outliers.
  - Illogical combinations of data
- Data structure and storage
  - Uniqueness
  - Referential integrity
  - Cardinality integrity
- Migration/integration
  - Normalization inconsistencies.
  - Duplicate or lost data



## Examples of What You Will Find

- A financial services company knew of 3 genders: M, F, and blank. They did not know about X and C.
- A home care products company discovered shipments slated for 16'x16' pallets. The IS manager wondered what kind of truck they would go on.
- Prior to a VA audit, a cross-check of medical billings by a healthcare provider showed it was performing open heart surgeries in ambulances.
- Consumer products mfr. learned a product of theirs was railroad boxcars.



## DQ Assessment Framework

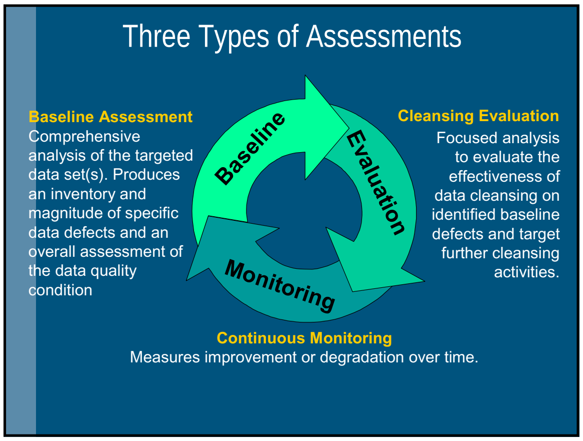
Meta Data

Assessment Preparation
Requirements
Define Scope
Define Business Rules

Assessment Measuring
Domain Analysis
Compl-eteness tests
Validity Tests
Struct-ural and Refer-ential Tests
Business Rule Tests

Findings Presentation
SME Analysis/ Verification
Findings Preparation
Management Reporting

Data



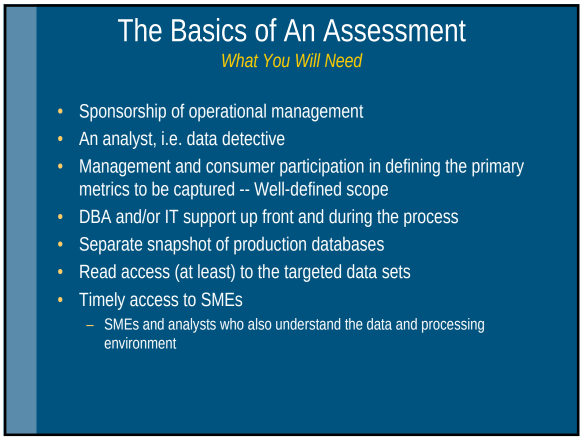
## Three Types of Assessments

**Baseline Assessment**
Comprehensive analysis of the targeted data set(s). Produces an inventory and magnitude of specific data defects and an overall assessment of the data quality condition

**Cleansing Evaluation**
Focused analysis to evaluate the effectiveness of data cleansing on identified baseline defects and target further cleansing activities.

Baseline
Evaluation
Monitoring

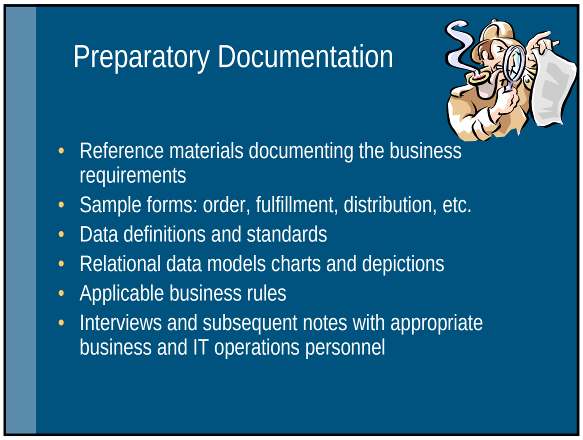
**Continuous Monitoring**
Measures improvement or degradation over time.



## The Basics of An Assessment

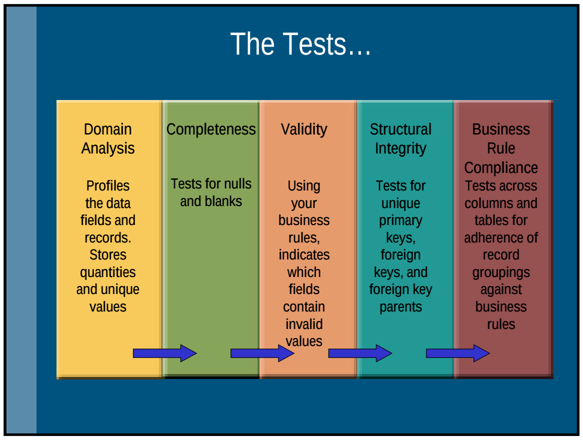
*What You Will Need*

- Sponsorship of operational management
- An analyst, i.e. data detective
- Management and consumer participation in defining the primary metrics to be captured -- Well-defined scope
- DBA and/or IT support up front and during the process
- Separate snapshot of production databases
- Read access (at least) to the targeted data sets
- Timely access to SMEs
  - SMEs and analysts who also understand the data and processing environment



## Preparatory Documentation

- Reference materials documenting the business requirements
- Sample forms: order, fulfillment, distribution, etc.
- Data definitions and standards
- Relational data models charts and depictions
- Applicable business rules
- Interviews and subsequent notes with appropriate business and IT operations personnel



## The Tests…

Domain Analysis

Profiles the data fields and records. Stores quantities and unique values

Completeness

Tests for nulls and blanks

Validity

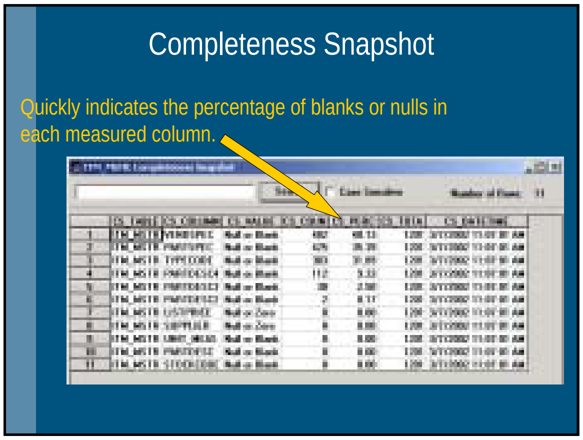
Using your business rules, indicates which fields contain invalid values

Structural Integrity

Tests for unique primary keys, foreign keys, and foreign key parents

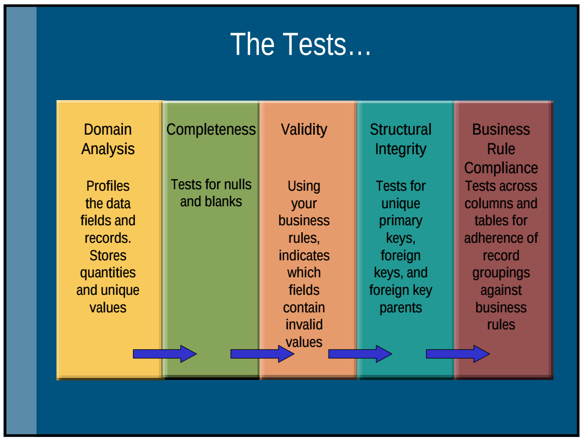
Business Rule Compliance

Tests across columns and tables for adherence of record groupings against business rules



## Completeness Snapshot

Quickly indicates the percentage of blanks or nulls in each measured column.



The Tests…

Domain Analysis

Profiles the data fields and records. Stores quantities and unique values

Completeness

Tests for nulls and blanks

Validity

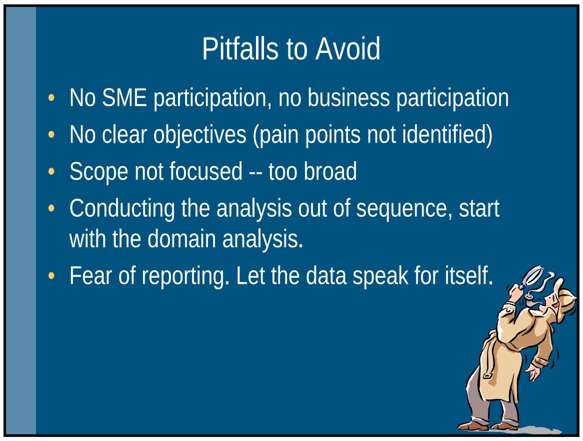
Using your business rules, indicates which fields contain invalid values

Structural Integrity

Tests for unique primary keys, foreign keys, and foreign key parents

Business Rule Compliance

Tests across columns and tables for adherence of record groupings against business rules



## Pitfalls to Avoid

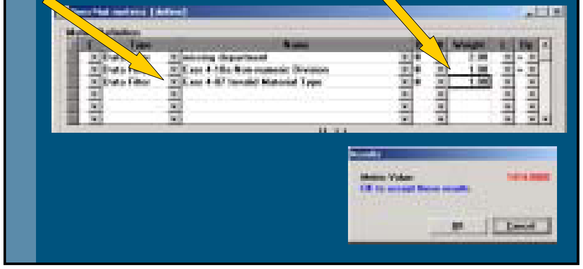
- No SME participation, no business participation
- No clear objectives (pain points not identified)
- Scope not focused -- too broad
- Conducting the analysis out of sequence, start with the domain analysis.
- Fear of reporting. Let the data speak for itself.

## Deliverables You Should Supply

- A assessment report containing:
  - Examples of specific defects
  - Anecdotes of impacts of the defects
  - Tabular/textual reports of domain measurements
  - Tabular/textual reports of analysis tests
  - Charts depicting metrics over time
  - Recommendation for process improvements
  - Recommendation of priority data elements  to cleanse

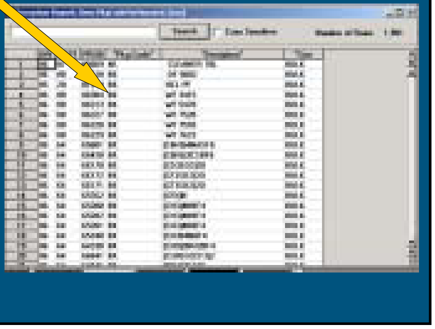
## Metrics?

Combines and/or compares multiple measurements with applied weighting.



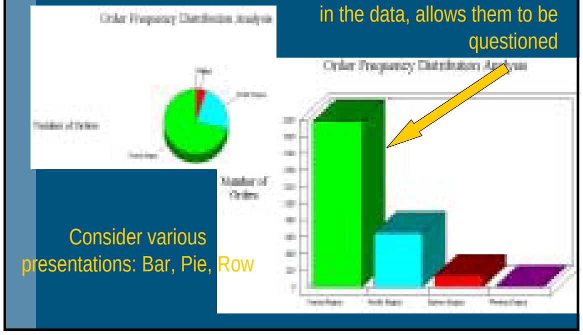
## Exception Reports

Shows all records with a non-numeric package code.



## Analysis Charts

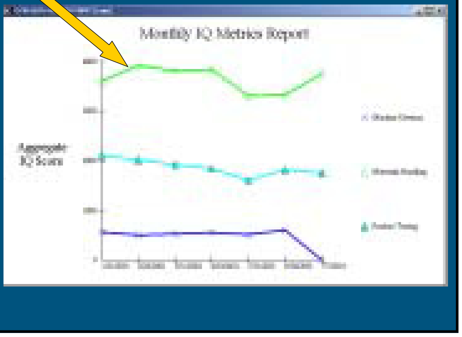
Shows distribution of values found in the data, allows them to be questioned



Consider various presentations: Bar, Pie, Row

## Trend Reports

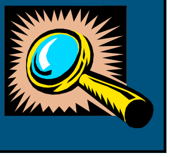
Can display trends of metrics or simple measurements.

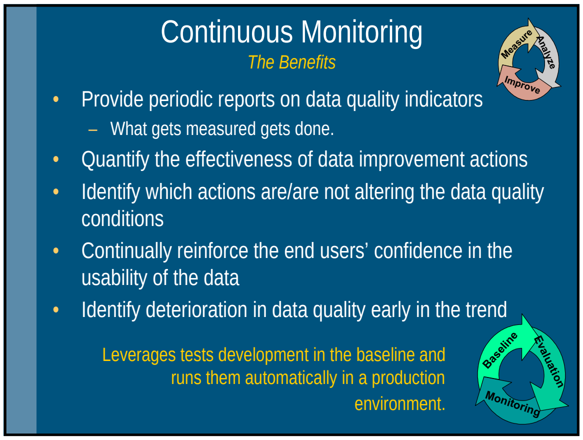


## Role of an Assessment Tool

*The Primary Functions*

- Data profiling and measurement
- Business rule auditing
- Problem identification, analysis and prioritization
- Meta data management, including history
- Trend analysis and continuous monitoring
- Analysis reporting and charting
- Data certification





## Continuous Monitoring

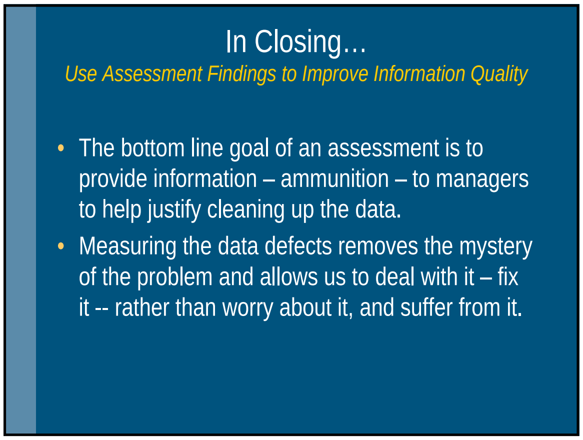
*The Benefits*

Measure
Analyze
Improve

- Provide periodic reports on data quality indicators
  - What gets measured gets done.
- Quantify the effectiveness of data improvement actions
- Identify which actions are/are not altering the data quality conditions
- Continually reinforce the end users' confidence in the usability of the data
- Identify deterioration in data quality early in the trend

Leverages tests development in the baseline and runs them automatically in a production environment.

Baseline
Evaluation
Monitoring



## In Closing…

*Use Assessment Findings to Improve Information Quality*

- The bottom line goal of an assessment is to provide information – ammunition – to managers to help justify cleaning up the data.
- Measuring the data defects removes the mystery of the problem and allows us to deal with it – fix it -- rather than worry about it, and suffer from it.

*7th International Conference on Information Quality (IQ 2002)*

## Questions