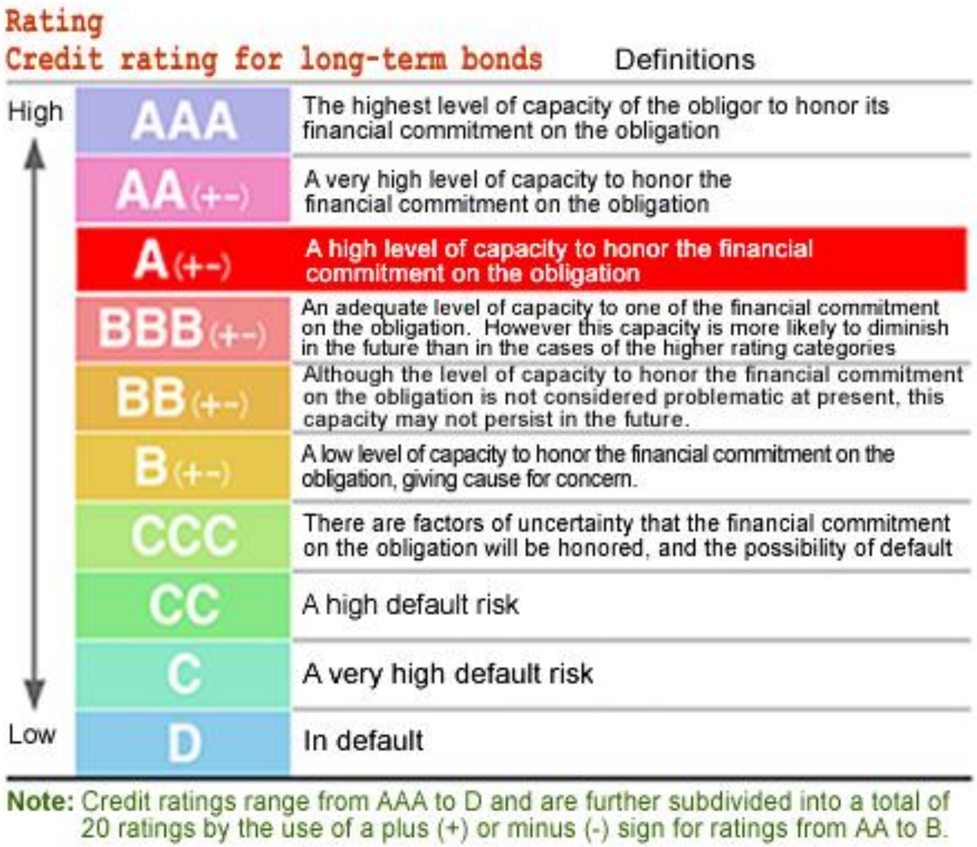


Rating

Credit rating for long-term bonds	Definitions
High AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
AA(+-)	A very high level of capacity to honor the financial commitment on the obligation
A(+-)	A high level of capacity to honor the financial commitment on the obligation
BBB(+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
BB(+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
B(+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
CC	A high default risk
C	A very high default risk
Low D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.

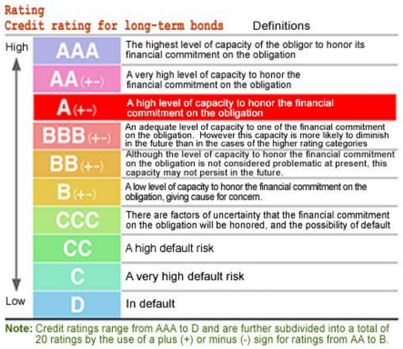
Minimizing Impact and Risk of Bad Data – Lessons from Other Industries

Sunil Gupta

Gupta Programming

Minimizing Impact and Risk of Bad Data

Analysis of data issues



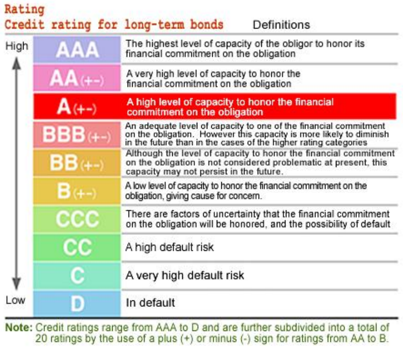
	Rating Credit rating for long-term bonds	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA (+-)	A very high level of capacity to honor the financial commitment on the obligation
	A (+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB (+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB (+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B (+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.

Data Issues

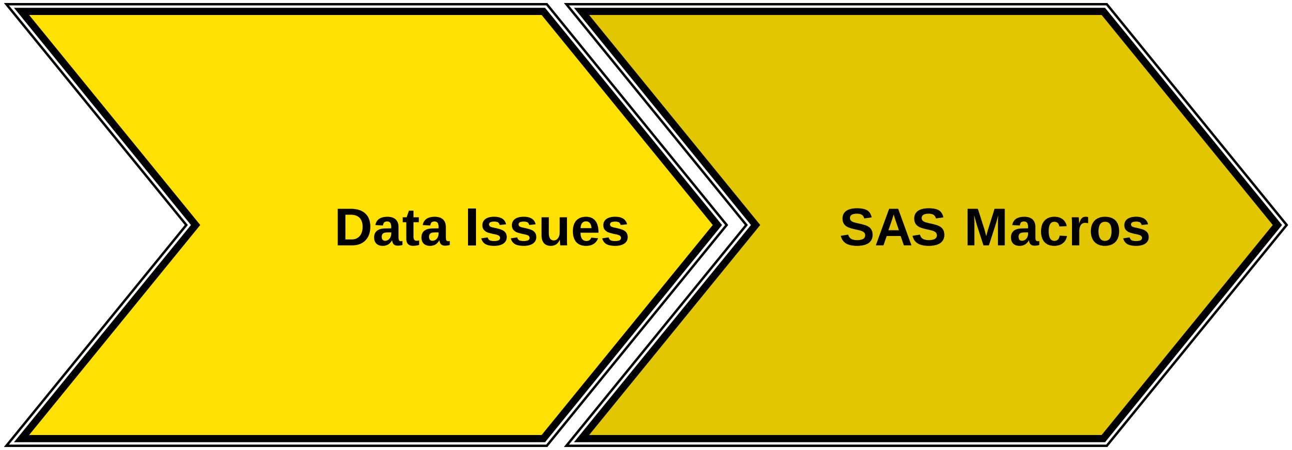
Minimizing Impact and Risk of Bad Data

Effective methods and SAS macros to identify data issues



	Rating Credit rating for long-term bonds	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA(+-)	A very high level of capacity to honor the financial commitment on the obligation
	A(+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB(+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB(+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B(+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.

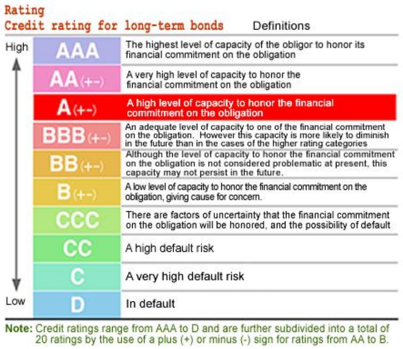


Data Issues

SAS Macros

Minimizing Impact and Risk of Bad Data

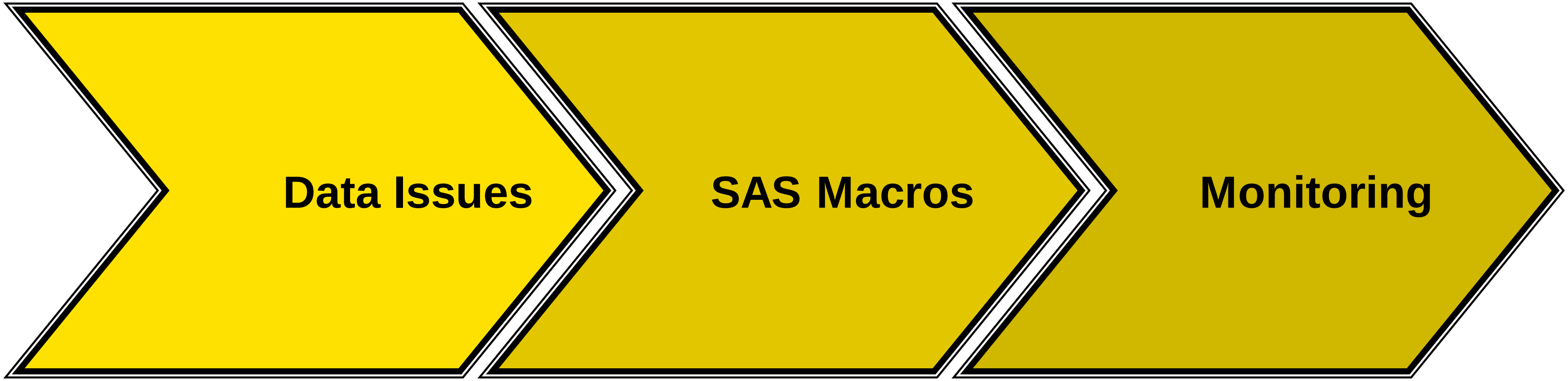
Communicating and monitoring improvements in data issues



Rating
Credit rating for long-term bonds

	Rating	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA(+-)	A very high level of capacity to honor the financial commitment on the obligation
	A(+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB(+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB(+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B(+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

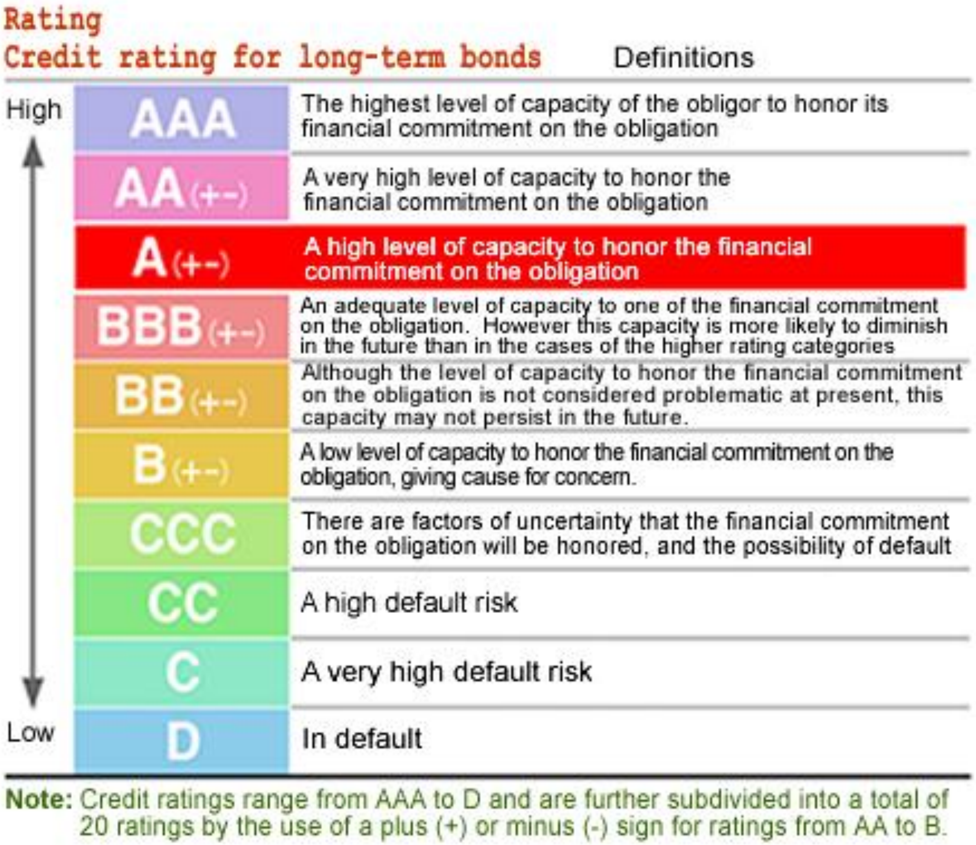
Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.



Data Issues

SAS Macros

Monitoring



Rating

	Credit rating for long-term bonds	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA (+-)	A very high level of capacity to honor the financial commitment on the obligation
	A (+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB (+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB (+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B (+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.

Analysis of data issues:

Identifying and quantifying the impact of bad data

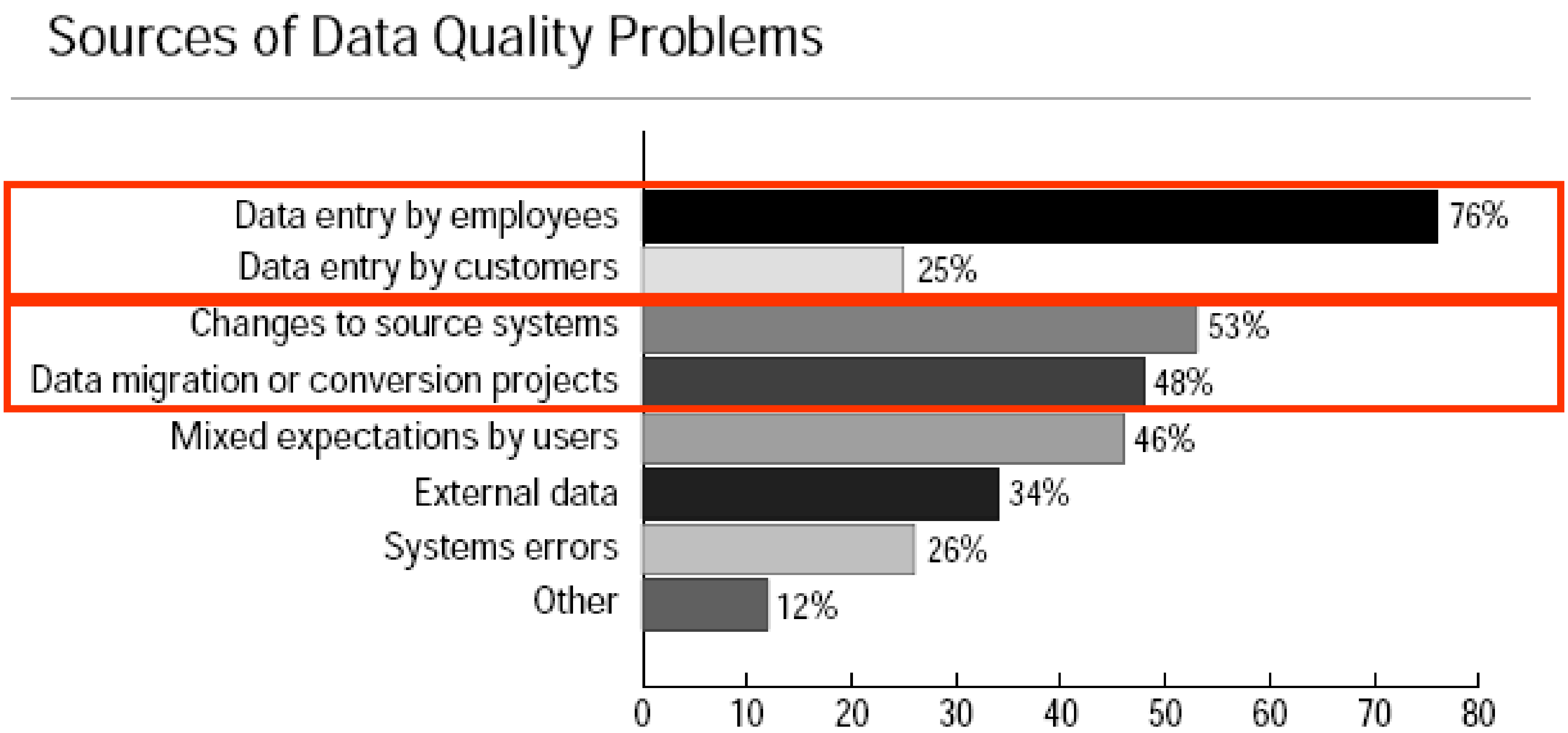
Bad data exists everywhere

- Duplicate records exist
- Missing values in required variables
- Start dates are after stop dates
- Invalid value for variable
- Poor-quality data vs. Fraud data (Trimming, Cooking, Altering, Forgery)?
- Breaking business rules: data better or worse than expected

Across all industries, bad data costs companies more than $ 600 billion per year.

For clinical trials, up to 5 % of raw data values may initially be erroneous.

Sources of Data Quality Problems across all Industries



2001 survey by the Data Warehousing Institute

Check Data: Each Data Transfer, Data Conversion or Data Updates

Most industries have a regulatory responsibility

- Incorrect or Incomplete clinical data
- Critical variables may need to be:
  - Non-missing
  - Consist only of valid values
  - Be within a range
  - Or be consistent with other variables.
- Process using Edit Check Macros
  - Unit Testing
  - Universal principles

21 CFR Part 11

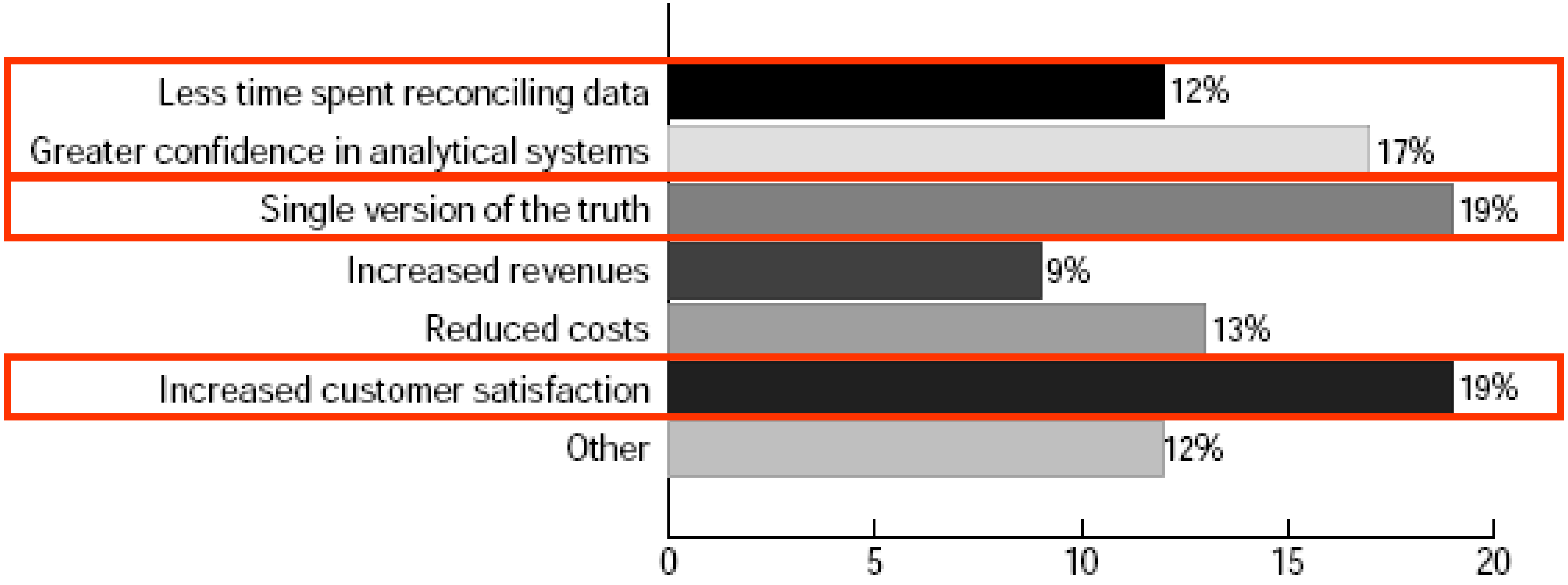
Prevent confusion and frustration.

Prevent incorrect clinical study conclusion: Safety, Efficacy

Comply with safety requirements

Benefits of High Quality Data across all Industries

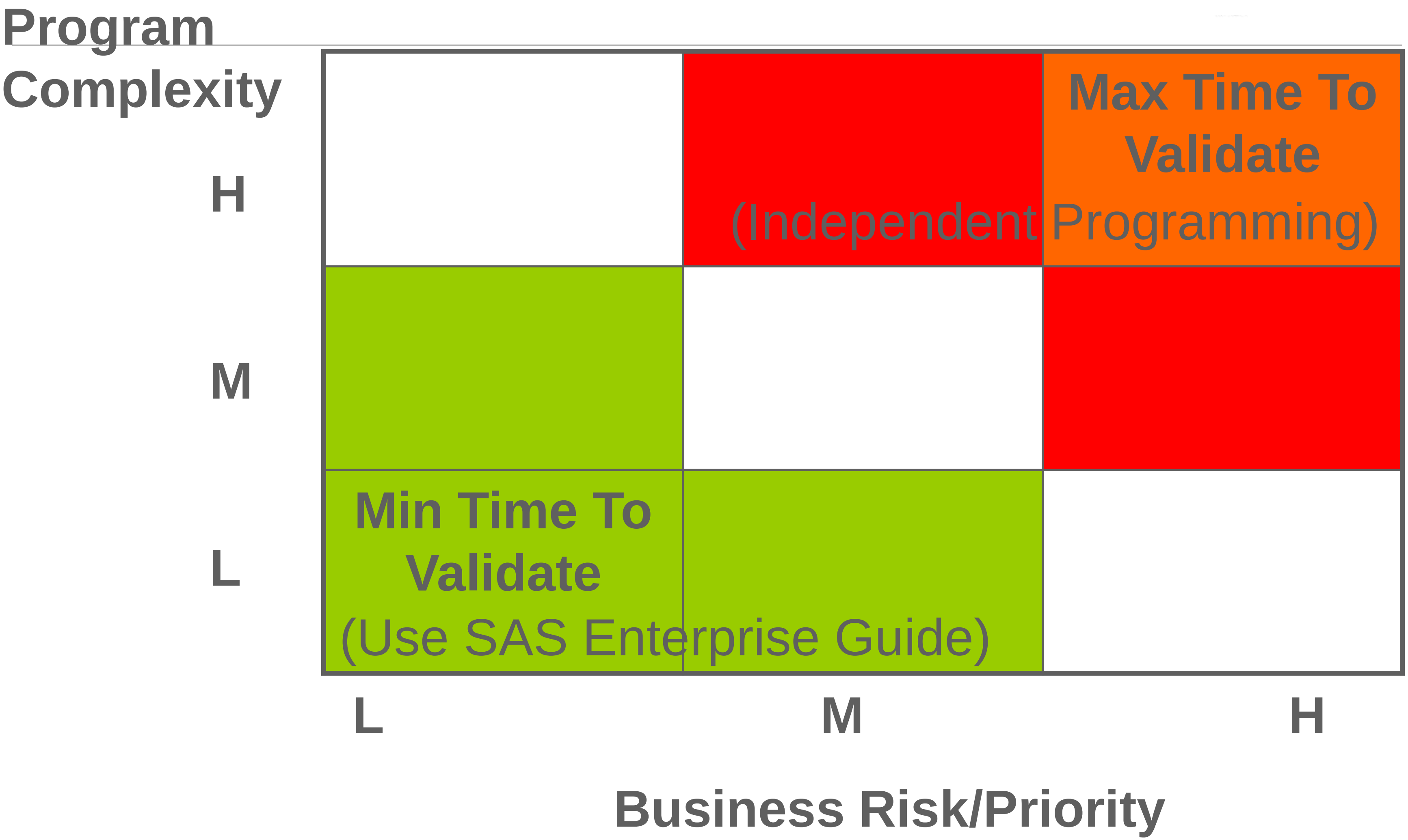
Benefits of High Quality Data



2001 survey by the Data Warehousing Institute

About $20 to $25 per case report form page or up to 15% of clinical research budget may be spent to ensure data quality.

One approach for Risk-Assessment and Validation



Program Complexity

H

M

L

Max Time To Validate

(Independent Programming)

Min Time To Validate

(Use SAS Enterprise Guide)

L

M

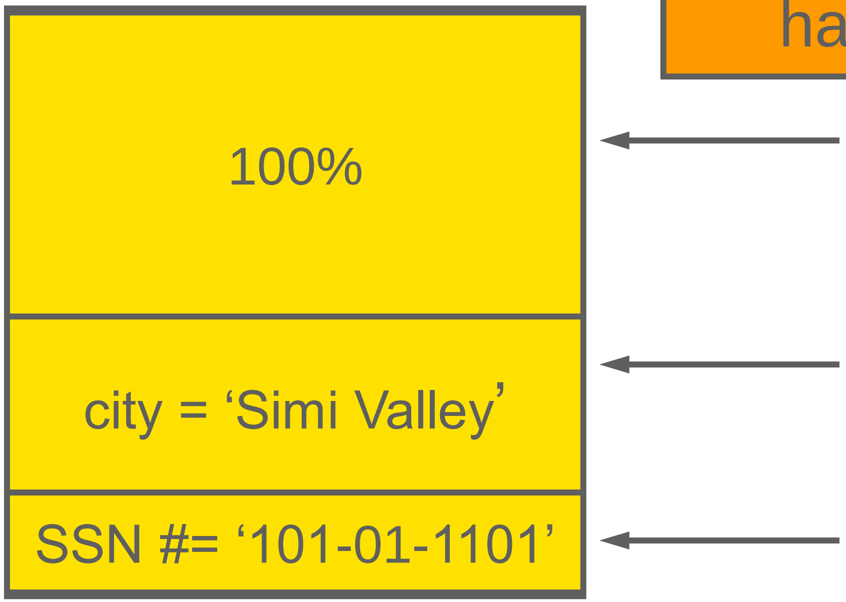
H

Business Risk/Priority

Track all datasets, tables, listings and figures. Do not underestimate the workload and complexity of processing financial data.

How much Data needs to be Validated?

Which validation methods have worked for you?



100%

city = ‘Simi Valley’

SSN #= ‘101-01-1101’

All data in data set (most time consuming)

All data for sample subset (focused or random)

All data for selected customer (first or problem customer)

<u>Adaptive Strategies as an alternative to 100% validation</u>

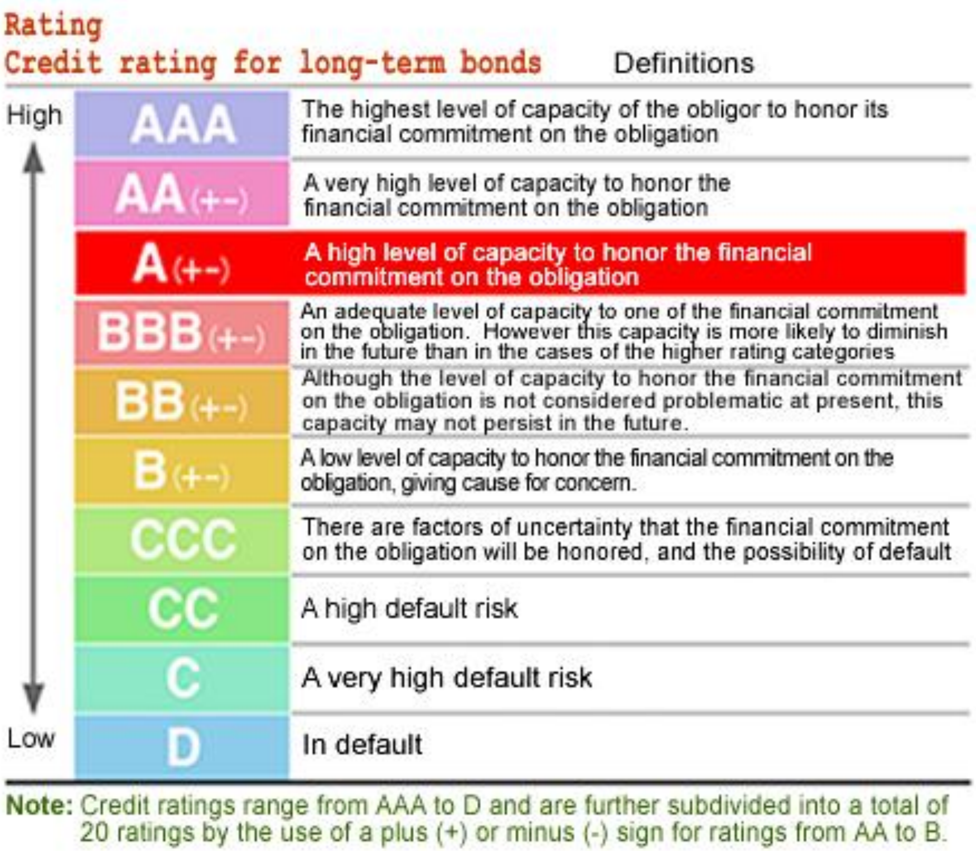
1. Start at 70% validation and increase or decrease percentage based on qc issues found from the first clinical study.
2. Validate based on risk category: High (90%), Median (80%), or Low (70%).

For Data Issues in Tables: Prevent False Positive and Negative Findings

	True Issue	False Issue
Issue Reported	Correct!	False Positive
Issue Ignored	False Negative	Correct!

Both False Positive and Negative can case fatal problems.

Analysis of data issues:



	Rating Credit rating for long-term bonds	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA(+-)	A very high level of capacity to honor the financial commitment on the obligation
	A(+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB(+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB(+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B(+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.

Establishing protocol criteria for clinical data acceptance testing example

Understand the Clinical Data Process Flow

Raw Data

Demog:	Valid/Invalid Data
Vitals:	Valid/Invalid Data
Labs:	Valid/Invalid Data
AE:	Valid/Invalid Data

As a result, in general:

- Keep invalid data in data sets

- Exclude invalid data in reports

Edit Check Process

1. Identify Invalid Data based on DMP
2. Isolate Data Issue
3. Communicate finding to CDM

Outcome

1. MONTHLY: Monitor Improvements in Invalid Data
   with e-mail notification to all team members
2. FINAL: Use Valid Data in Analysis data sets, Tables,
   Lists and Graphs

Apply Data Acceptance Testing

- Create Data Management Plan

- Generally a monthly process to refresh data
- Identify, isolate and report clinical data issues
- Make critical decision before database lock to accept or reject database

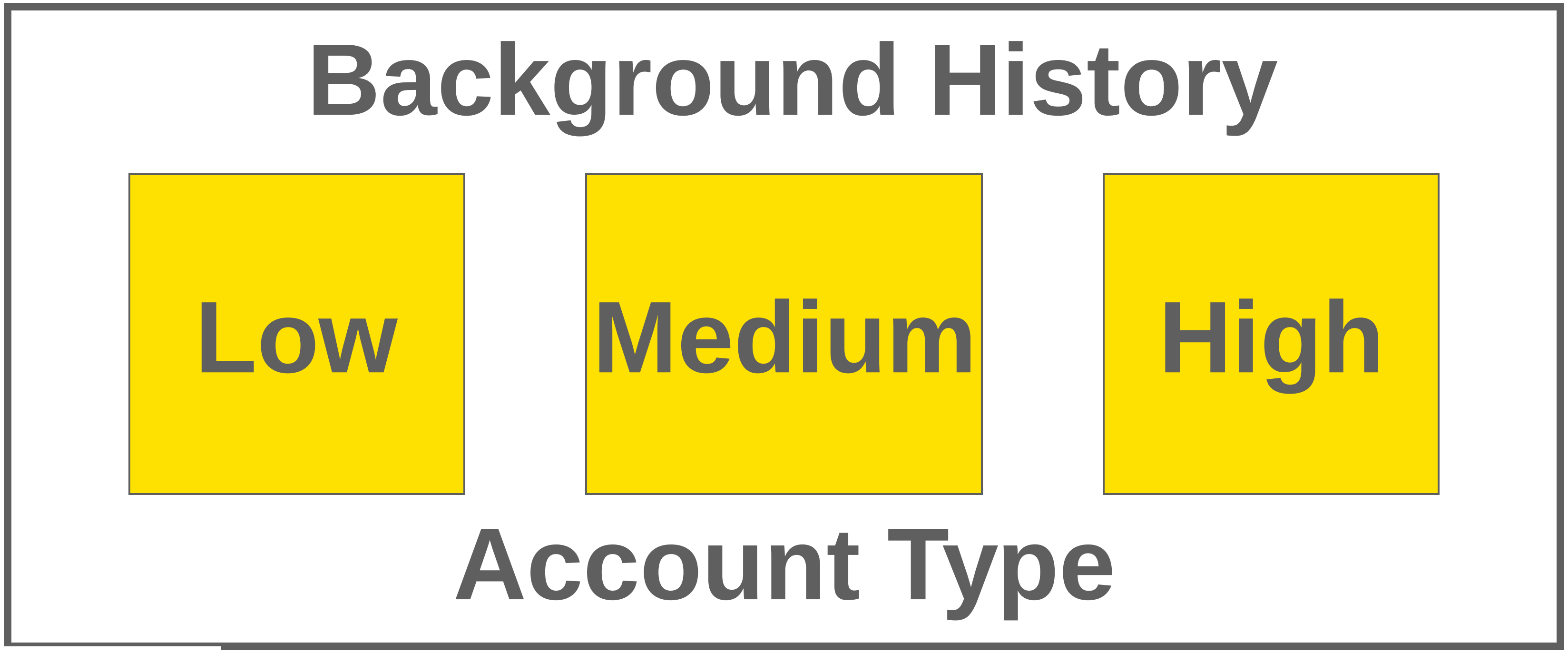
Similar to Data Acceptance Testing for software application.

Check Data Compliance at all three levels

I. Background History checks

II. Risk Profile checks

III. Account Type checks



Background History

Low

Medium

High

Account Type

Solution involves four steps **before** Database Lock

- 1. Specifying Requirements in Data Management Plan (DMP)
- 2. Developing and Testing Edit Check Macros
- 3. Communicating Results with Clinical Data Management (CDM) (Partnership)
- 4. Monitoring the Metrics of Data Issues

Using Edit Check Macros standardized our approach to validate the quality of raw clinical data.

Specifying Requirements in Data Management Plan (DMP)

1. All unique key variables are required.
   - Patient ID variable is non-missing and unique.

2. Confirm minimum and maximum values.
   - Vitals data set: valid temperature and blood pressure values within lower and upper range values.

3. Display all unique values of selected variables.
   - Demog data set: valid treatment (active, placebo).

continued …

At a minimum, these types of data checks should be performed.

Specifying Requirements in Data Management Plan (DMP)

4. Confirm the logic between two variables.
   - Adverse Events data set: adverse event description, preferred term and system organ class need to be consistent.

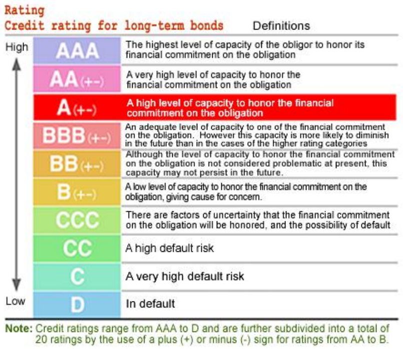
5. Confirm the consistency between two clinical dates.
   - Adverse Events data set: Adverse start dates before or same day as adverse stop dates.

6. Are patient follow-up visit windows in compliance with the protocol?

At a minimum, these types of data checks should be performed.

Minimizing Impact and Risk of Bad Data

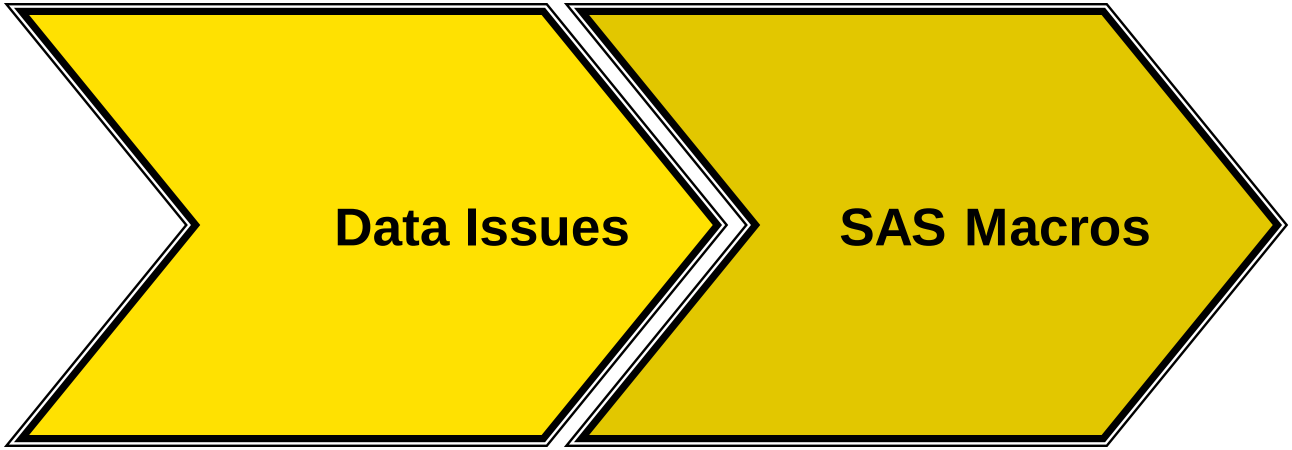
Effective methods and SAS macros to identify data issues



Rating

	Credit rating for long-term bonds	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA(+-)	A very high level of capacity to honor the financial commitment on the obligation
	A(+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB(+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB(+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B(+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.



Data Issues

SAS Macros

Effective methods and SAS macros to identify data issues:

Developing and testing edit check macros

Develop and Test Edit Check Macros

- System Requirements
  - Unselect data checks
  - Easily modify data checks, Add new data checks
  - Display ‘No records found’ for no data issues
  - Display feedback from CDM on data issues
- Limited Programming Resources
  - SAS’s ODS, Minimum SAS macro programming
  - Simple, task-oriented macros approach
  - Apply standard options to selected SAS Procedures

A traditionally lengthy SAS program of over 1,000 lines is easier to read with only 75 lines containing 75 edit check macro calls.

continued …

Confirm Edit Check Macros: Functional Requirements

- Macros use basic macro programming techniques that are easy to understand
  - Quick development of new macros
  - Quick enhancements of existing macros

- Macros provide informative feedback in titles
  - Input data set name
  - Variables checked
  - Any subset condition applied

Data set name, variables checked and subset condition are important information for CDM to first confirm the data issue before taking any action.

continued …

Confirm Edit Check Macros: Functional Requirements

- Macros provide reference information in footnotes
  - Program name
  - Output file name
  - Date executed

- Macros display data issues
  - Patient and visit identification
  - Data values of variable checked
  - Supporting variables (if any)
  - One data issue/page
  - Findings saved to one RTF file

Ideally Edit Check Macros have the right Balance

Right amount of balance between calling standard macros and programming capability.

1. 80% of standard programming can be done in 20% of the time with edit check macros.
2. Leveraging your knowledge with SAS programming and data.
3. Self documenting since macro call contains all relevant information.
4. More compact code to see more tasks in limited window size.

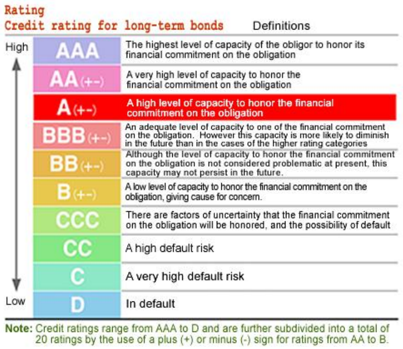
Design Strategy: One Edit Check Macro for each type of data issue (partial list)

Type of Data Issue	Brief Description
Acceptable Values	Values are one of the valid values for variable
Consistency Across Variables	Values are consistent across multiple variables
Consistency Across Data sets*	Values are consistent across multiple data sets
Non-Duplicate Records	Each record is unique and not duplicated
Required Value	Value is non-missing

* May require extra programming step since most all edit check macros require single data set.

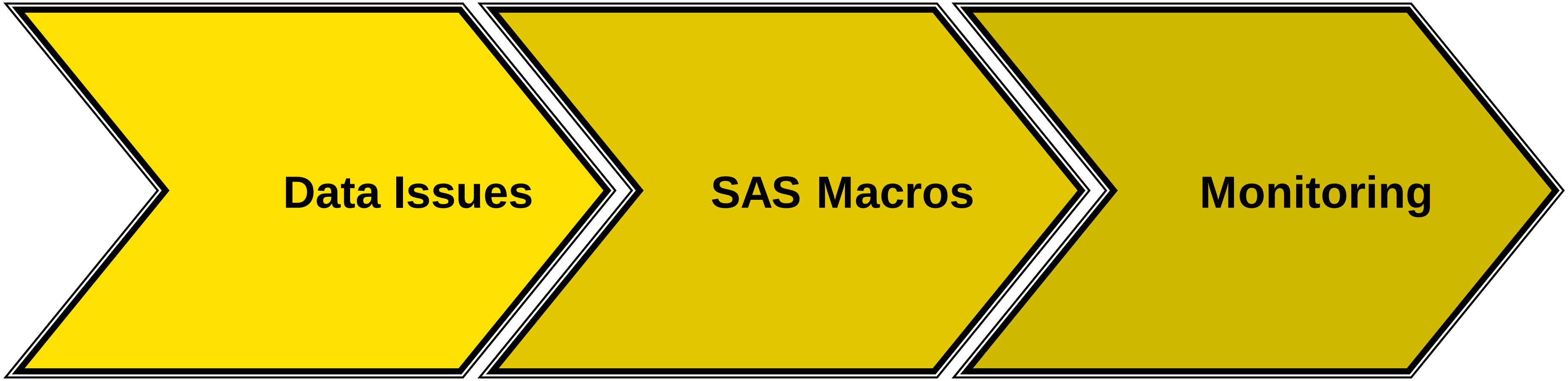
Minimizing Impact and Risk of Bad Data

Communicating and monitoring improvements in data issues



	Rating Credit rating for long-term bonds	Definitions
High	AAA	The highest level of capacity of the obligor to honor its financial commitment on the obligation
	AA(+-)	A very high level of capacity to honor the financial commitment on the obligation
	A(+-)	A high level of capacity to honor the financial commitment on the obligation
	BBB(+-)	An adequate level of capacity to one of the financial commitment on the obligation. However this capacity is more likely to diminish in the future than in the cases of the higher rating categories
	BB(+-)	Although the level of capacity to honor the financial commitment on the obligation is not considered problematic at present, this capacity may not persist in the future.
	B(+-)	A low level of capacity to honor the financial commitment on the obligation, giving cause for concern.
	CCC	There are factors of uncertainty that the financial commitment on the obligation will be honored, and the possibility of default
	CC	A high default risk
	C	A very high default risk
Low	D	In default

Note: Credit ratings range from AAA to D and are further subdivided into a total of 20 ratings by the use of a plus (+) or minus (-) sign for ratings from AA to B.



Data Issues

SAS Macros

Monitoring

Managing metrics on data acceptance testing for quicker decisions

Establish Metrics on Data Acceptance Testing

- Unit: # of edit checks tested (based on DMP)

- Summary level measurement (Scope of issue – at least one failed patient out of x edit checks)

  = # of failed edit checks/Total # of edit checks

- Detail level measurement (Impact of issue – systematic or localized problem based on # of patients affected?)
  - Overall = # of failed records/(Total # of records x Total # of edit checks)
  - By edit check = # of failed records/Total # of records

Goal: To capture and monitor the correction of unexpected data.

Results of Simple Example: 1 data set, 5 checks, 10 records

Checks	Fail	Pass	% Failed
1. Work History	1	9	10%
2. Credit Cards	2	8	20%
3. Tax Returns	2	8	20%
4. Loans	1	9	10%
5. Checking/401(k) Account	0	10	0%
Total	**6**	**44**	**12%**

Monitor and Communicate Metrics on Data Acceptance Testing

- Unit: 5 edit checks tested

- Summary level measurement (Scope)

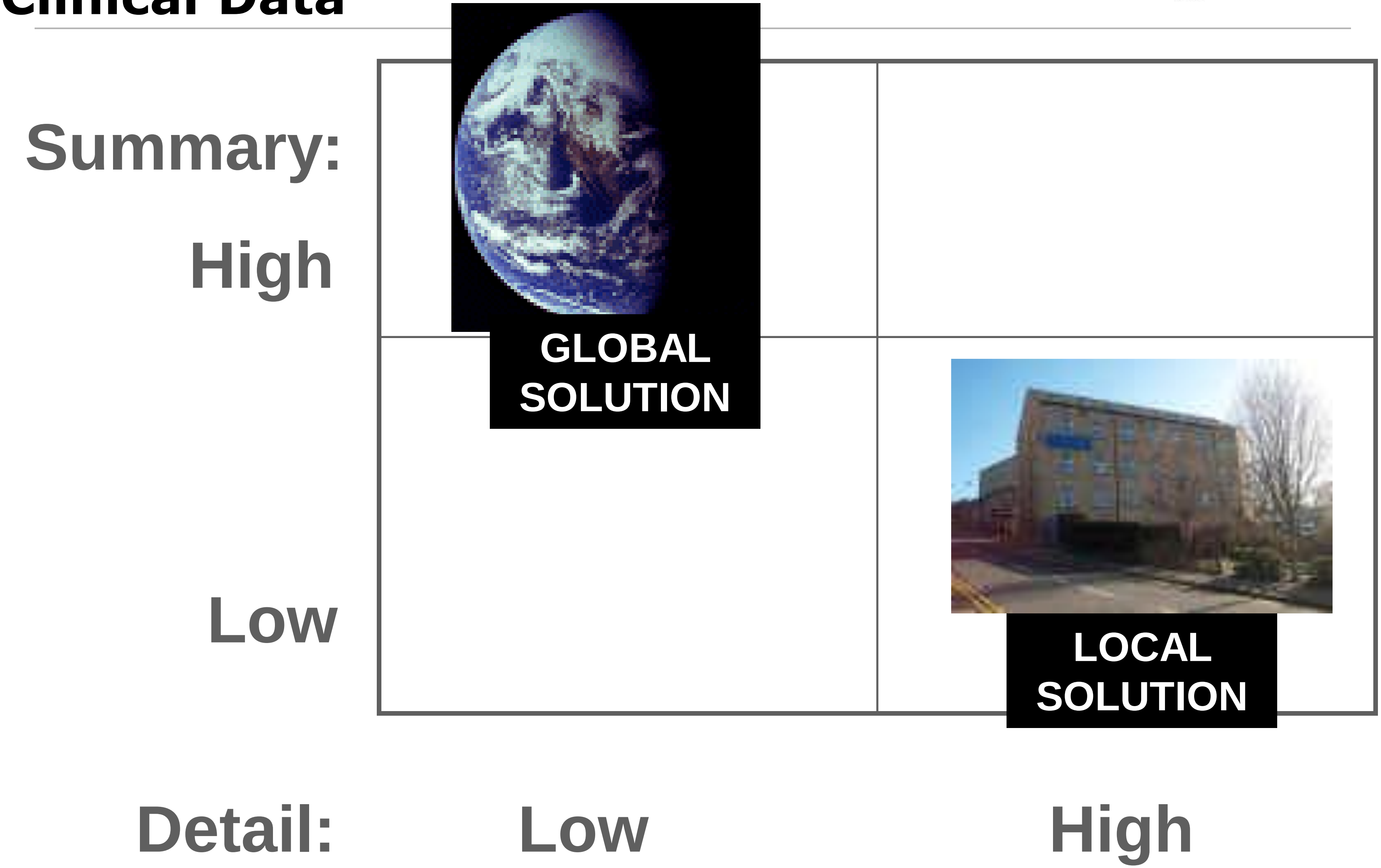
  80% = 4 failed edit checks/5 edit checks

- Detail level measurement (Impact)
  - Overall 12% = 6 failed records/(10 records x 5 edit checks)
  - By edit check: 1. Work (10%), 2. Credit Cards (20%), 3. Tax (20%), 4. Loans (10%), 5. Accounts (0%)

Analyze Possible Reasons for Poor Finance Data

Summary: High	Global Scope Ex. 80% Summary, 12% Detail - incorrect age calculation	
Low		Local Scope Ex. 20% Summary, 20% Detail - poor site
Detail:	Low	High

Analyze Possible Reasons for Poor Clinical Data



Summary:

High

Low

GLOBAL SOLUTION

LOCAL SOLUTION

Detail: Low High

Minimizing Impact and Risk of Bad Data – Lessons from Other Industries

Sunil Gupta

www.SASSavvy.com
www.GuptaProgramming.com
Sunil@GuptaProgramming.com