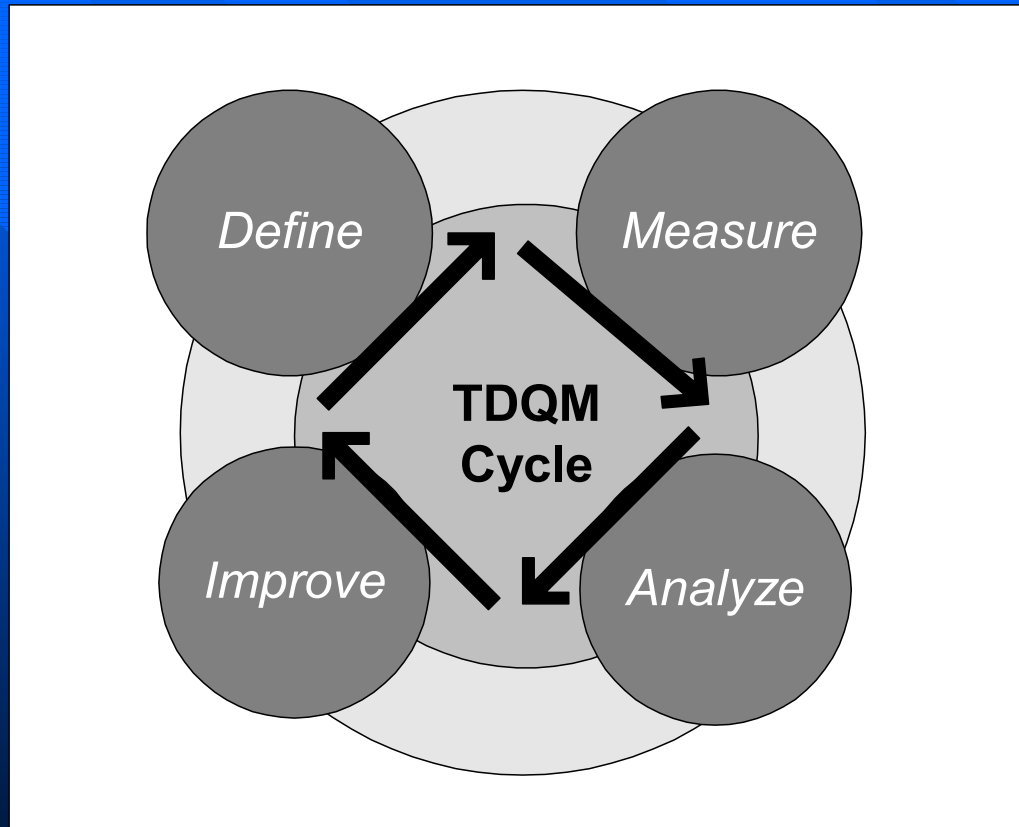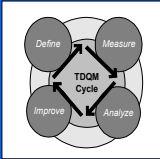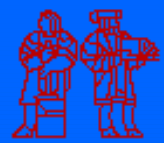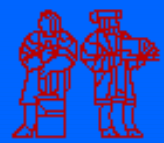# Introduction

- Total Data Quality Management (TDQM) Program

- Co-directed by Professors Stuart Madnick at MIT, and Richard Wang at Boston University

- Decade of Research and Practice

- Annual International Conference on Information Quality since 1996

- http://WEB.MIT.EDU/TDQM/

- e-mail: rwang@bu.edu

- Tel: (617) 739-7234, (617) 353-2678
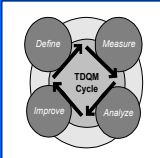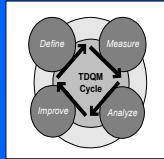
Total Data Quality Management Cycle

Product vs. Information Manufacturing
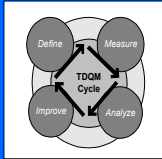
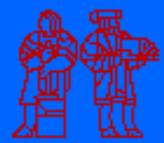|	**Product Manufacturing**	**Information Manufacturing**
Input	Raw Materials	Raw Data
Process	Assembly Line	Information System
Output	Physical Products	Information products
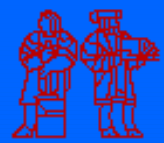
What is Data Quality?

■ Conventional View:

Data Quality = Accuracy

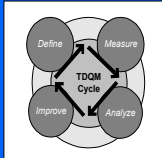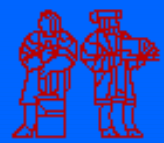# What is Data Quality?

■ Conventional View:

  Data Quality = Accuracy

■ New Research Finding

  ???

Data Quality is Multi-Dimensional

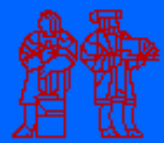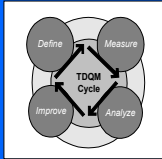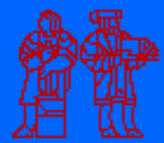| DQ Category	DQ Dimensions
Intrinsic **DQ**	Accuracy, Objectivity, Believability, Reputation
Accessibility **DQ**	Access, Security
Contextual **DQ**	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational **DQ**	Interpretability, Ease of understanding, Concise representation, Consistent representation, Ease of Manipulation
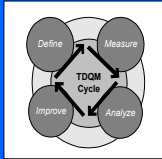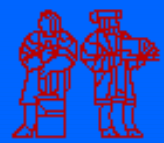
Why Is Data Quality Important?

- ■ Data of high quality is a valuable asset
- ■ Data of high quality can increase customer satisfaction
- ■ Data of high quality can improve revenues and profits

- ■ For a major west coast hospital
  - ● Fundamental to Clinical Value Improvement Process (CVIP)
  - ● Support strategic Business Development initiatives
  - ● Meet JCAHO ORYX Reporting Requirements
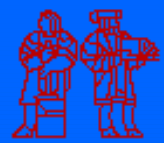 - ● Avoid OIG Audit due to data discrepancies
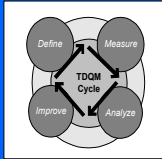
The Journey to Data Quality

- Adopt a customer perspective
- Adopt and adapt classical TQM principles
- **Manage data as a product**
  - Understand the consumer's data needs
  - Manage data as the product of a well-defined production process
  - Manage data as a product with a life cycle
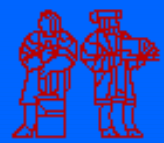 - Appoint a data product manager to manage the data product.

How to manage data as a product

- Identify a data product based on the consumer's needs

- Cluster related data products as a product line

- Develop data product maps that encompass
  - Data Producers, Custodians, and Consumers
  - Data Flow Network
  - Boundaries where Data are transferred from one group to another

- Develop data quality metrics

- Assign Data Ownership, Accountability, and Reward
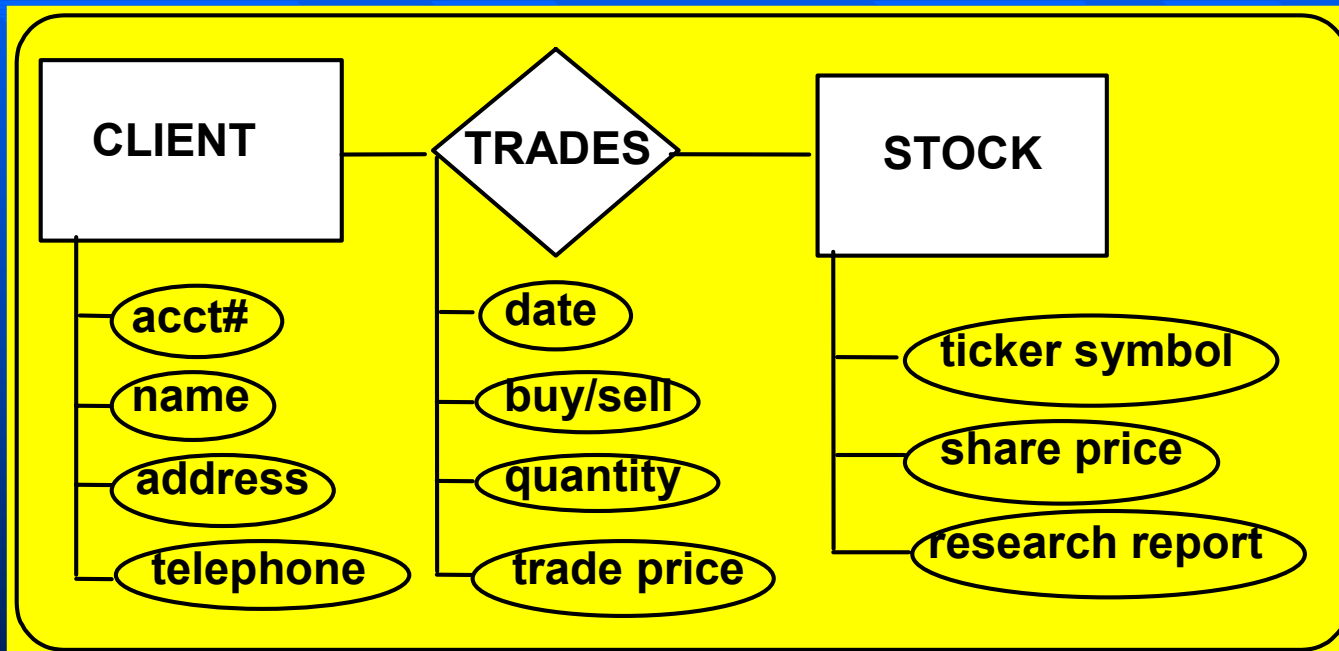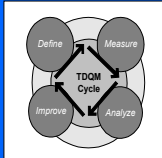
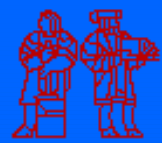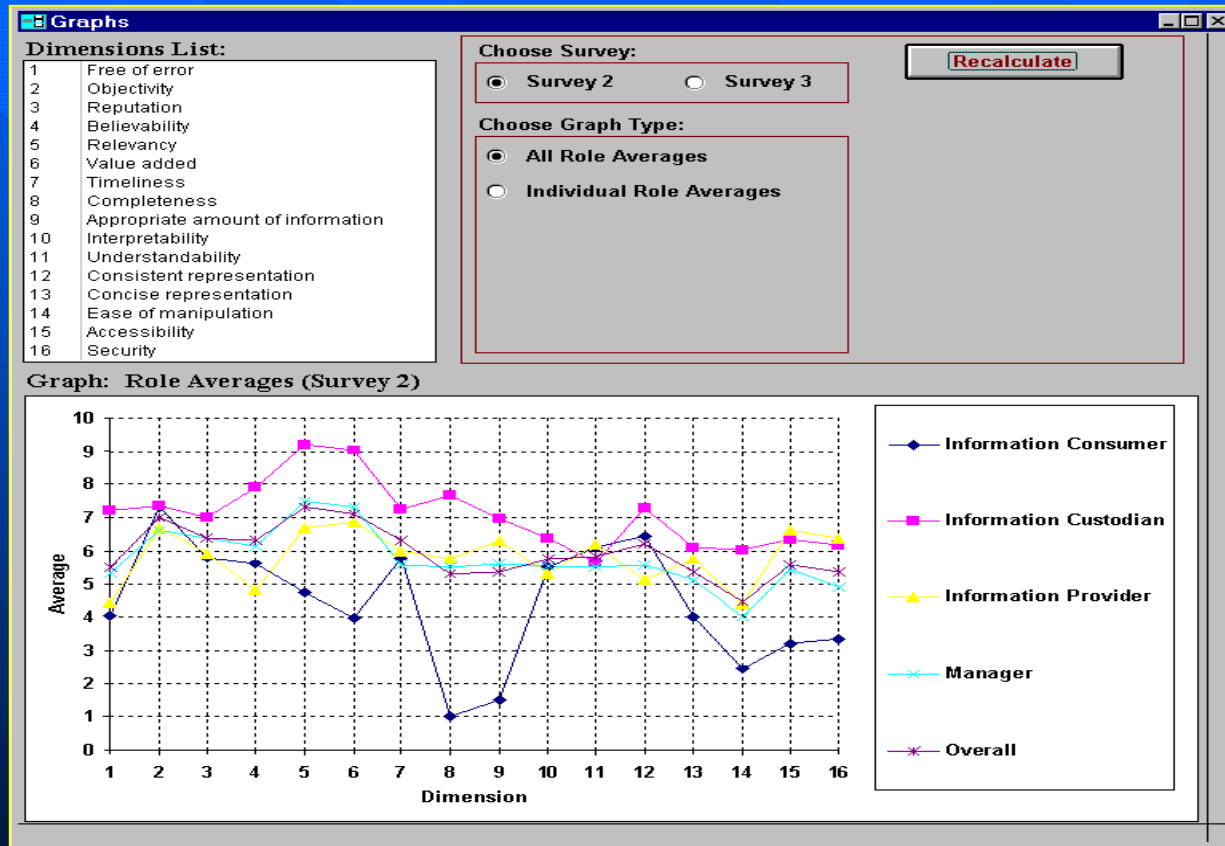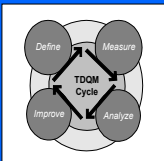- Continuous Improvement in the data product life cycle
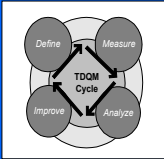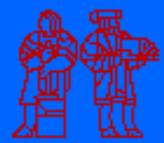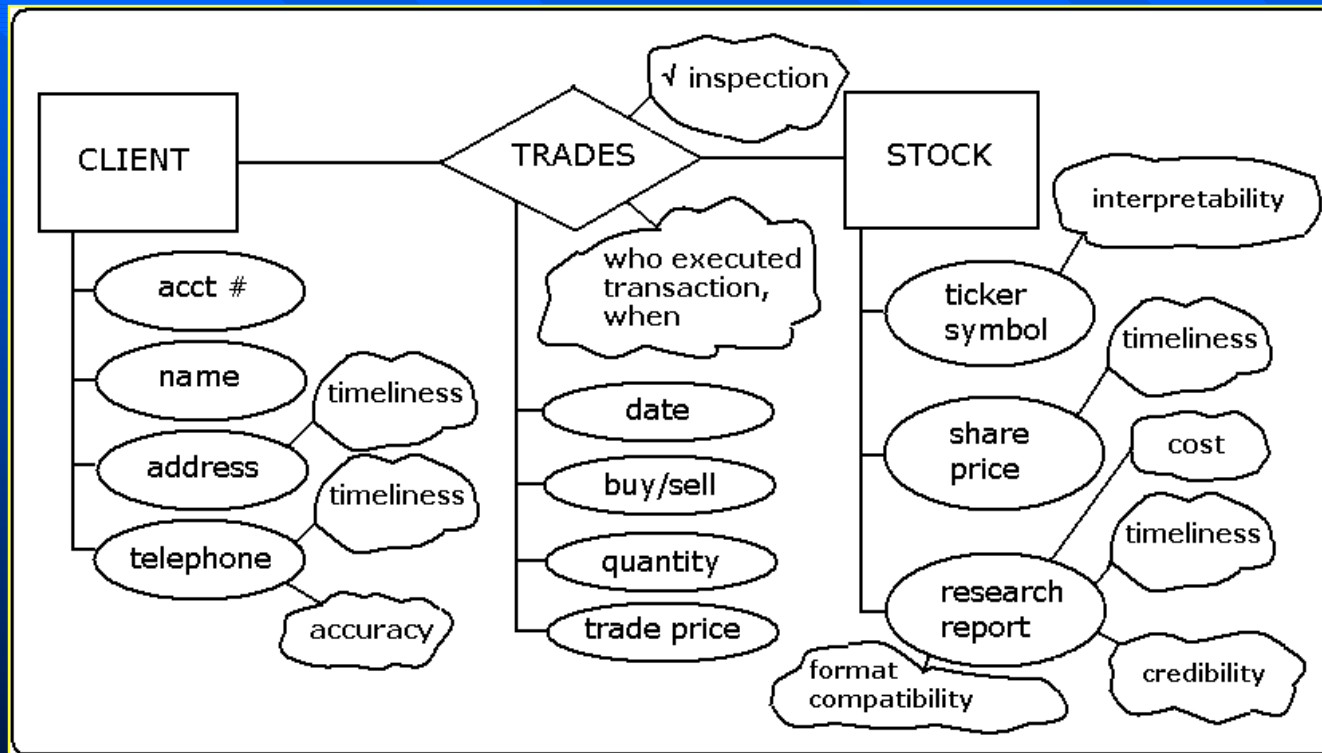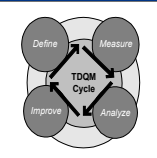
A Schematic of the TDQM Framework
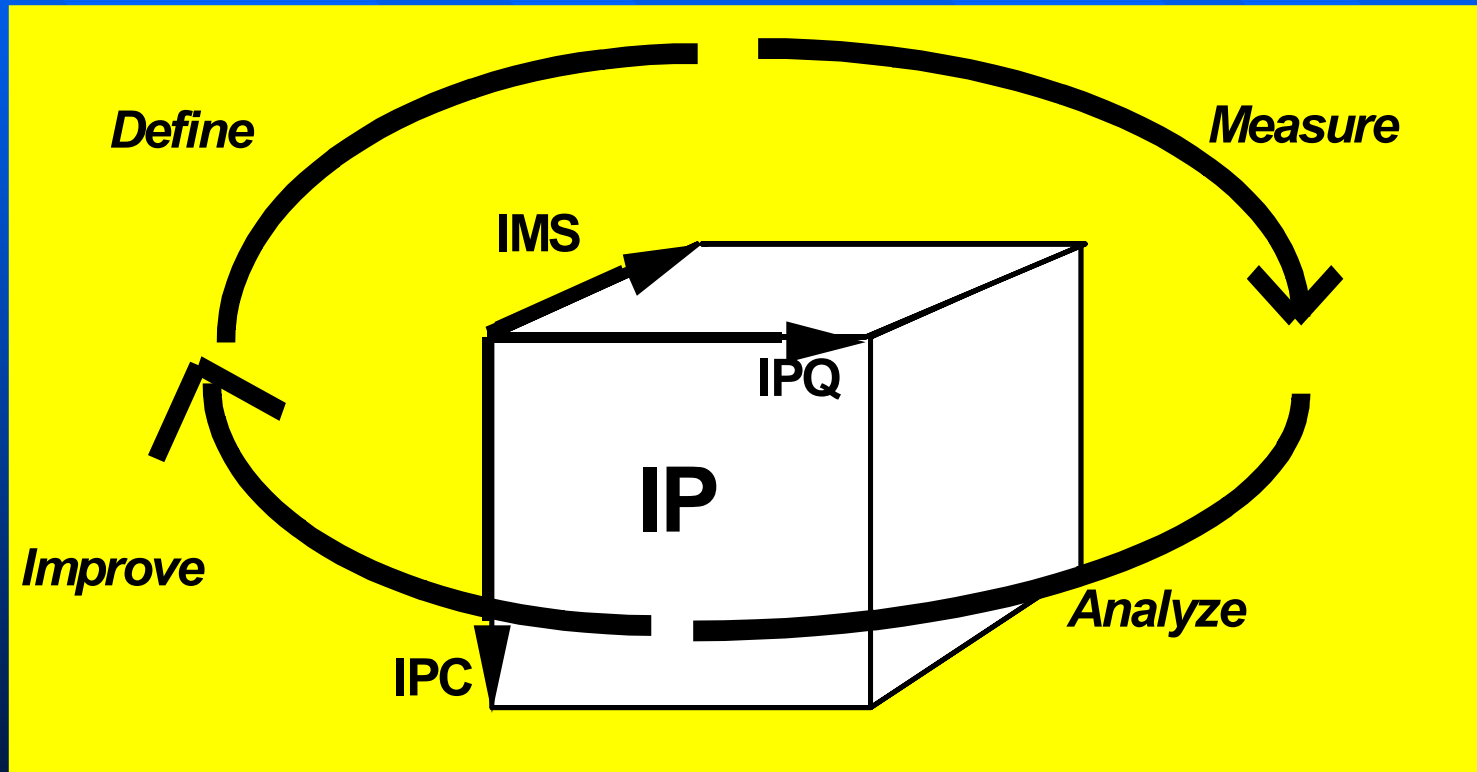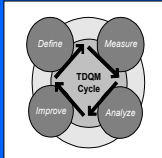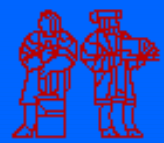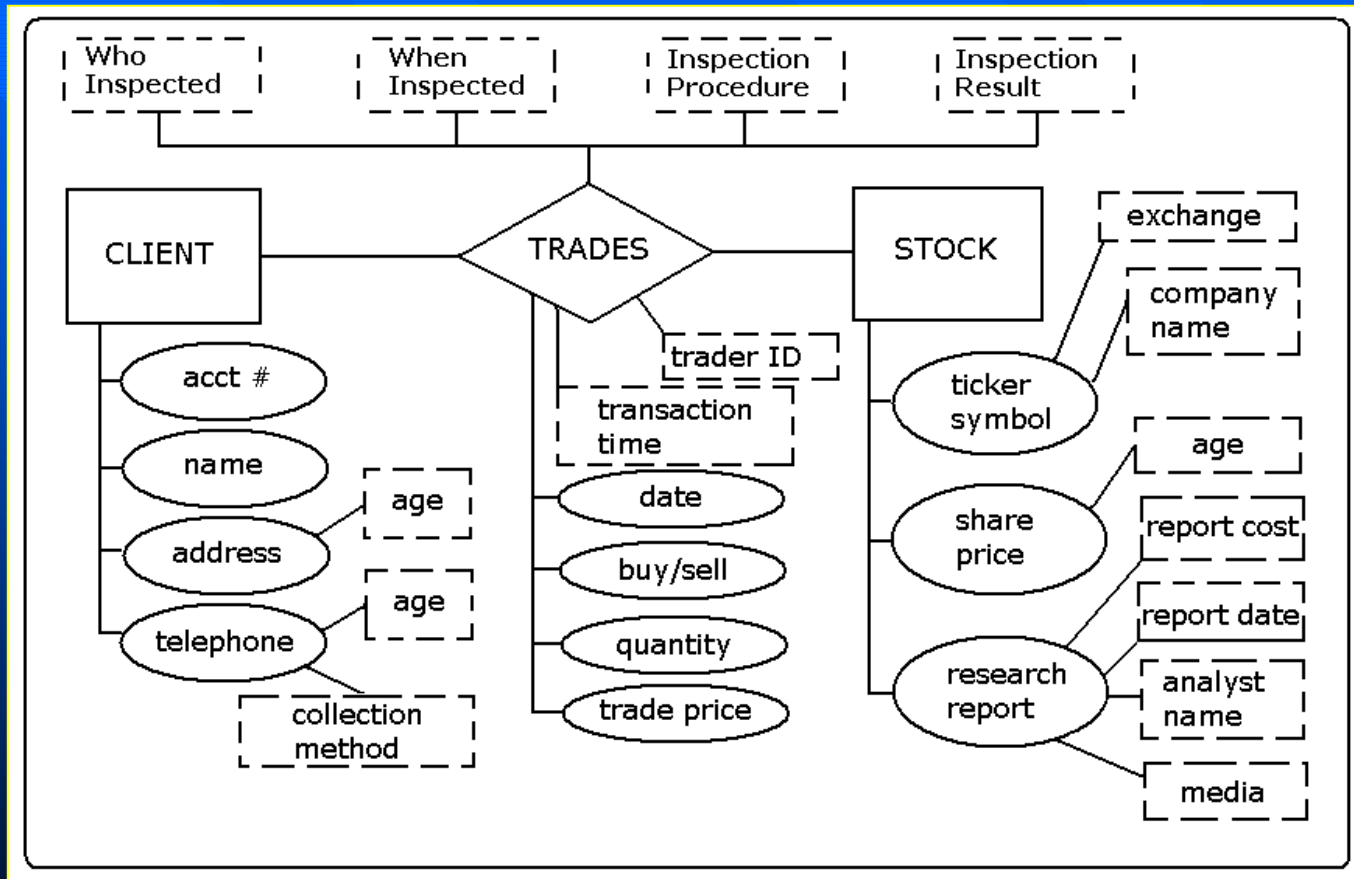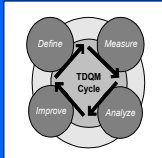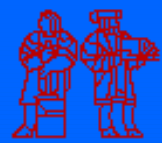
Can IP conform to specifications?

Dimensional Assessment of IQ Importance

IQ Added to the ER Diagram

A QER Diagram

An Illustrative IMS
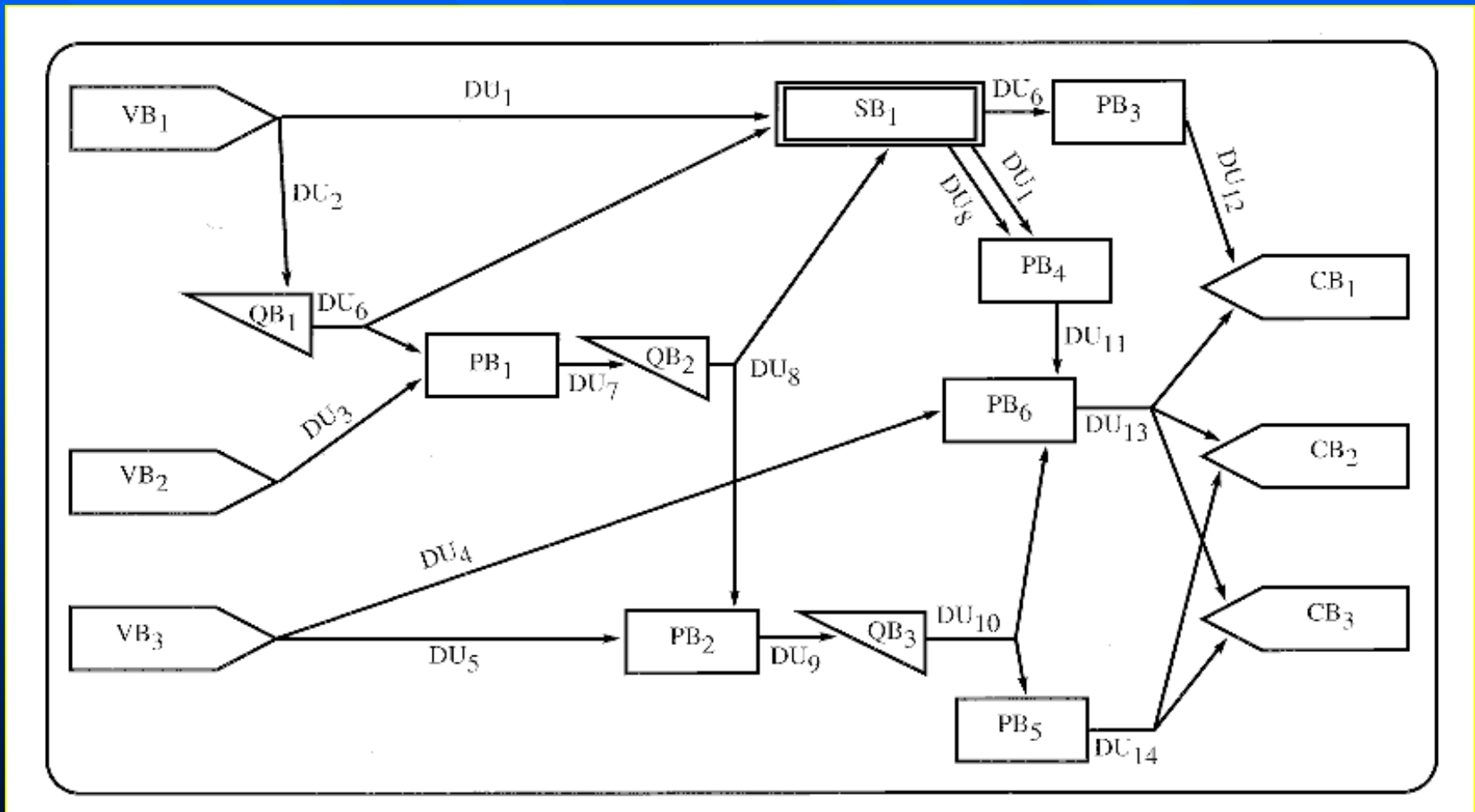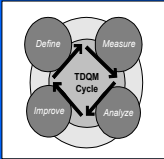
Concluding Remarks

- ■ Presented Key Concepts underlying the decade TDQM work
  - ● TDQM Cycle adapted from Deming Cycle
  - ● Analogy of product manufacturing goes a long way
  - ● Data Quality is multi-dimensional and goes beyond accuracy
  - ● It's a *journey*, stupid!
- ■ Database Theoretic Approach
  - ● Extending ER to QER
  - ● Extending the Relational Model to capture DQ attributes
  - ● Others in Data Quality (Kluwer Academic, 2000)

Concluding Remarks, Continued

■ Ostensibly missing is TQM/Statistical approach

- Not my forte

- Covered to a great extent by Goldberg, English and Redman

- Data Quality Control by Liepins & Uppuluri

■ TDQM Research Directions

- Journey to Data Quality: A Roadmap

- Information Supply Chain Management

- Information Product Map in Action
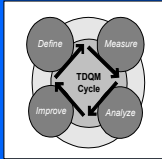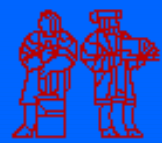
About the Presenter

Richard Y. Wang is Associate Professor at Boston University and Co-Director for the Total Data Quality Management program at Massachusetts Institute of Technology, where he had been a professor for a decade. Professor Wang has published extensively in top journals to develop concepts, principles, tools, methods, and techniques related to data quality. He has also published books on data quality including *Data Quality (*Kluwer Academic, 2000*), Quality Information and Knowledge* (Prentice Hall, 1999), *Data Quality Systems* (Cambridge Market Intelligence, 1995), and *Information Technology in Action* (Prentice Hall). Professor Wang received his Ph.D. degree from M.I.T. with a concentration in information technology.