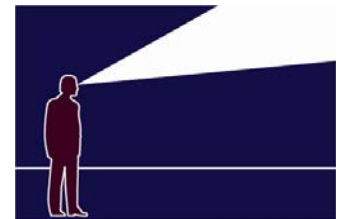


Data Quality 101: for IAIDQ Book Club

Thomas C. Redman, Ph.D.
Navesink Consulting Group

April 2006

www.dataqualitysolutions.com





Objectives

Provide an overview of “second-generation data quality systems.”

Provide a short summary of the main ideas.

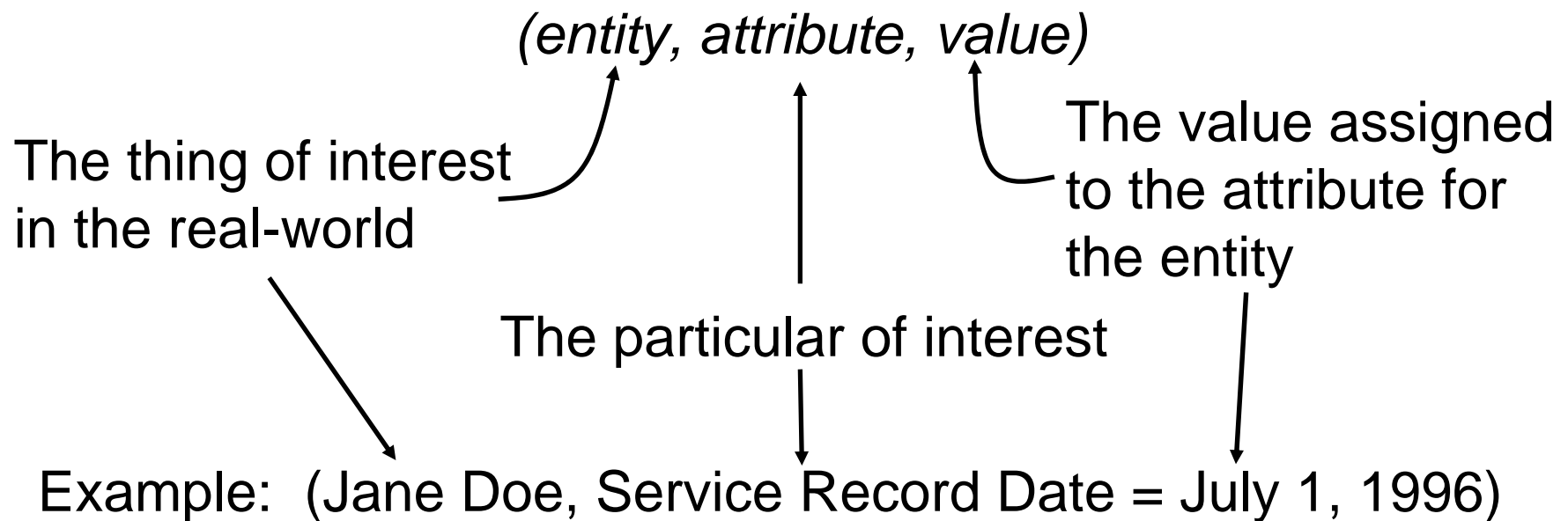
Answer some basic questions:

- What are data? What is data quality?
- What are the competing approaches to data quality management?
- How do those with the best data do it?

Show how the components of a second-generation data quality system work together and reinforce one and other.

Data - Defined

A *datum* consists of three elements:



Note that, as defined, data are abstract. "Customers" see them as they are *presented* in tables, databases, graphs, etc.



Implications...

Thus even the simplest datum arises from three distinct sources:

- The *model* (entity, attribute)-pair is created within a modeling process, usually by IT or purchased from outside.
- The data *value* is created (at enormous rates) by “the business.”
- The *presentation* may be created by database tools, application programs, PowerPoint presenters, etc.

This point has enormous implications for managing data quality.

Data Quality

Data are of high quality if they are fit for their intended uses (by customers) in operations, decision-making, and planning (after Juran).

Data that's fit for use

```
graph TD; A["Data that's fit for use"] --> B["free of defects:"]; A --> C["possess desired features:"]; B --> B1["- accessible"]; B --> B2["- accurate"]; B --> B3["- timely"]; B --> B4["- complete"]; B --> B5["- consistent with other sources"]; B --> B6["- etc."]; C --> C1["- relevant"]; C --> C2["- comprehensive"]; C --> C3["- proper level of detail"]; C --> C4["- easy-to-read"]; C --> C5["- easy-to-interpret"]; C --> C6["- etc."];
```

free of defects:

- accessible
- accurate
- timely
- complete
- consistent with other sources
- etc.

possess desired features:

- relevant
- comprehensive
- proper level of detail
- easy-to-read
- easy-to-interpret
- etc.

Customers are the ultimate arbiters of quality!!

There are quite literally, hundreds of “dimensions of data quality”

But experience suggests a few most basic needs:

- Data customers need to be able to *find and access* what they need (technology).
- They want the data to be *up-to-date and accurate* (values).
- They need to be able *understand what the data mean* (model).
- They want the data presented to them in ways that are *easy-to-understand* (presentation).
- For data they don't create themselves, they want to be able to get *help* when they need it.
- In large companies, they want the data to be *consistent* across various sources.
- *Privacy and security* are becoming more and more important!



Data Quality Systems

An organization's *data quality system* (even if only implicitly) is the totality of an organization's effort devoted to data quality, including efforts to:

- ❑ Find and correct errors.
- ❑ Remedy the impact of poor quality data.
- ❑ Understand customer needs.
- ❑ Measure performance against those needs.
- ❑ Close the gaps.
- ❑ Manage sources of data.
- ❑ Understand the impact of poor quality data.
- ❑ Build high-quality data into new processes and computer systems.
- ❑ Etc.

A Database is Like a Lake



Approaches to Data Quality: Defect Prevention

Most companies' current approach to data quality. Typical error rates are 1-5% and "cost of poor data quality" may be 20% of revenue.

FIRST-GENERATION:*
Inspection and Rework,
to find and fix defects

SECOND-GENERATION:*
Process/Supplier Management,
to prevent defects

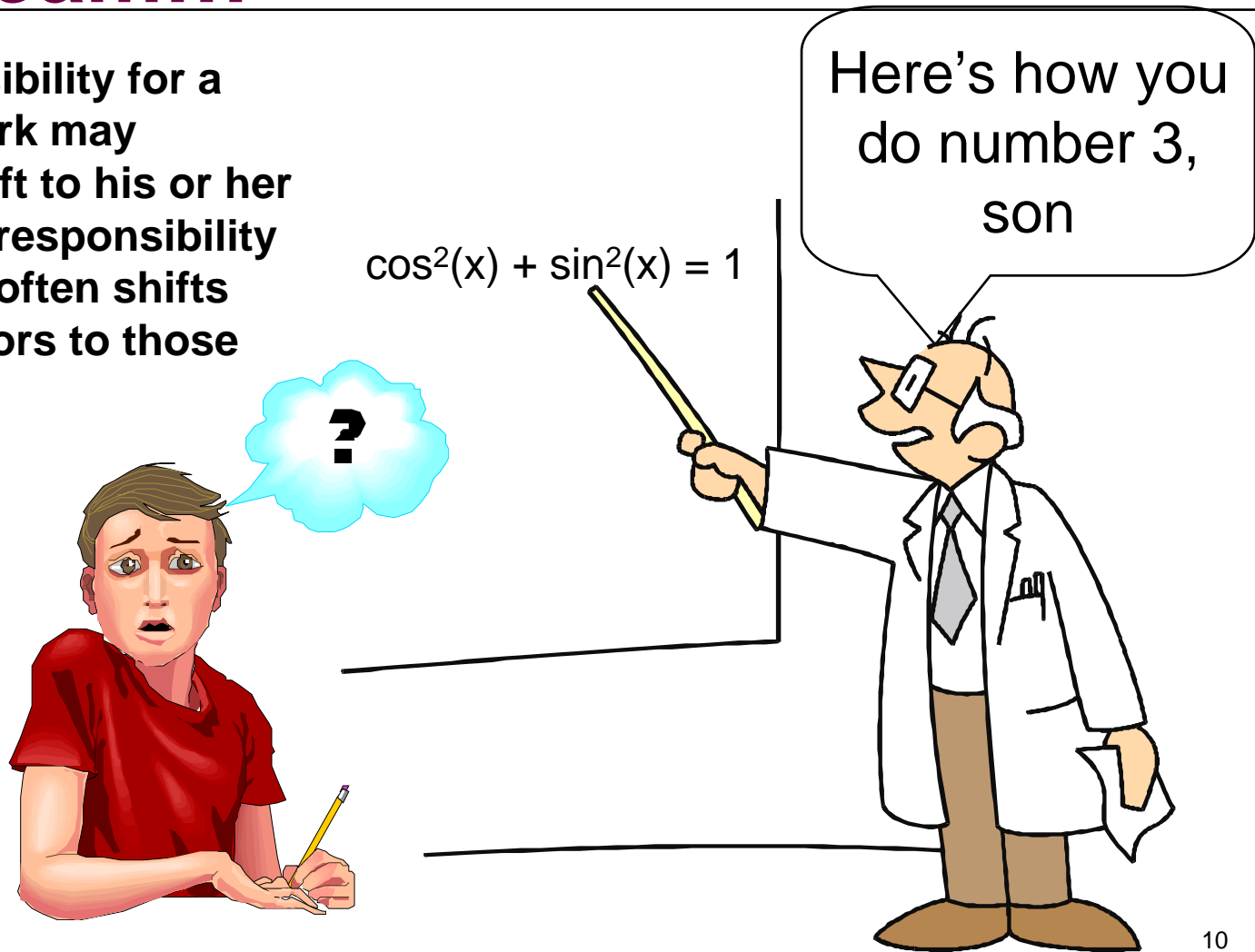
THIRD-GENERATION:*
Design,
defects "impossible"
*Don't know
of anyone here*

To accomplish this, original sources of data are held accountable. Typical error rates are 1-2 orders of magnitude better and the cost of poor data quality is reduced about two-thirds.

*terms after Ishikawa

It is so easy for accountability to shift downstream!!!

Just as responsibility for a child's homework may (magically?) shift to his or her parents, so too responsibility for data quality often shifts from their creators to those who use them.



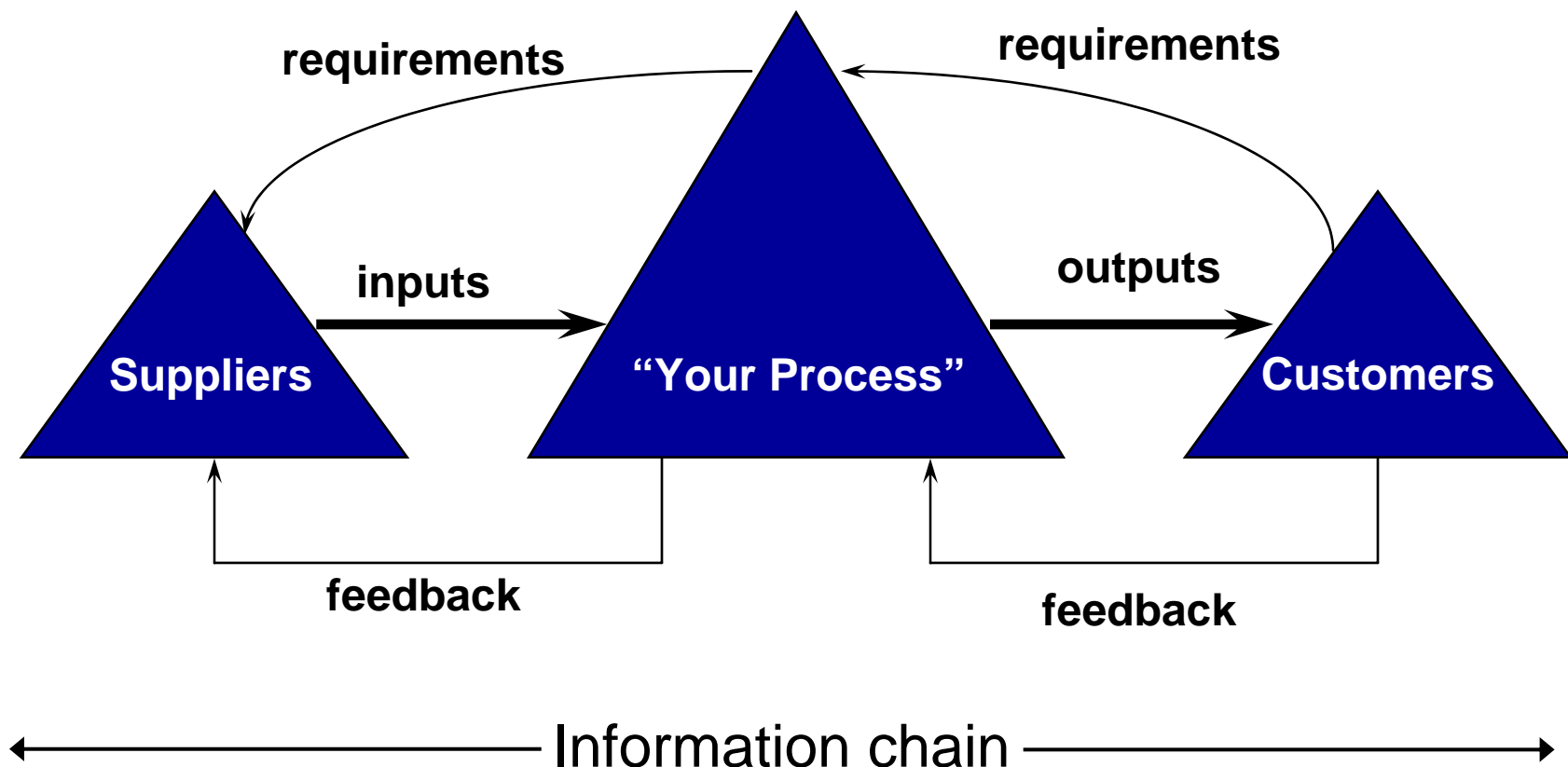
Second-Generation Data Quality Systems™

Those with the highest quality data assign management accountability for data at their original sources.

These original sources include:

- *External suppliers*, and they use *data supplier management* to obtain the best possible data from these suppliers.
- *Internal processes*, and they use *information chain management* to create the best possible data.
 - Everyday business (operations, decision-making)
 - Data modeling
 - “Standards-setting”
 - Application development
- In some cases, these responsibilities are codified in *policy*.
- Finally, *leadership* comes from very high levels.

Customer-Supplier Model



Second-Generation Data Quality Systems™

Those with the highest quality data focus on the data that serves “the most important needs of the most important customers:”

- Business issues/opportunities
- Customers and customer needs
- Data

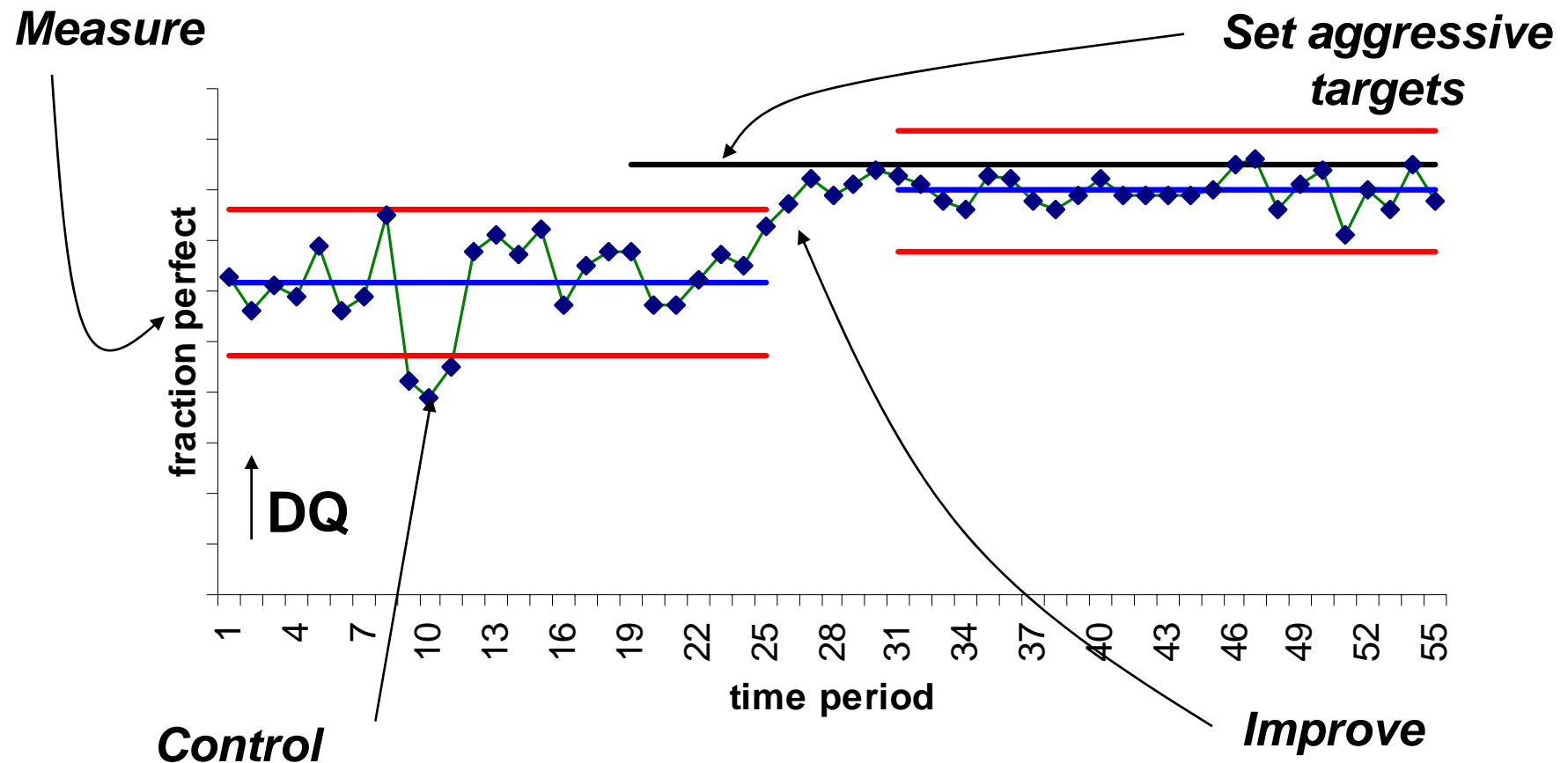
Those with the highest quality data focus on the most important “dimensions:”

- Accessibility
- Accuracy
- Clear Definition
- Consistency

NOTE: 50% of data are never used by anyone for anything

Second-Generation Data Quality Systems™

Those with the highest quality data:

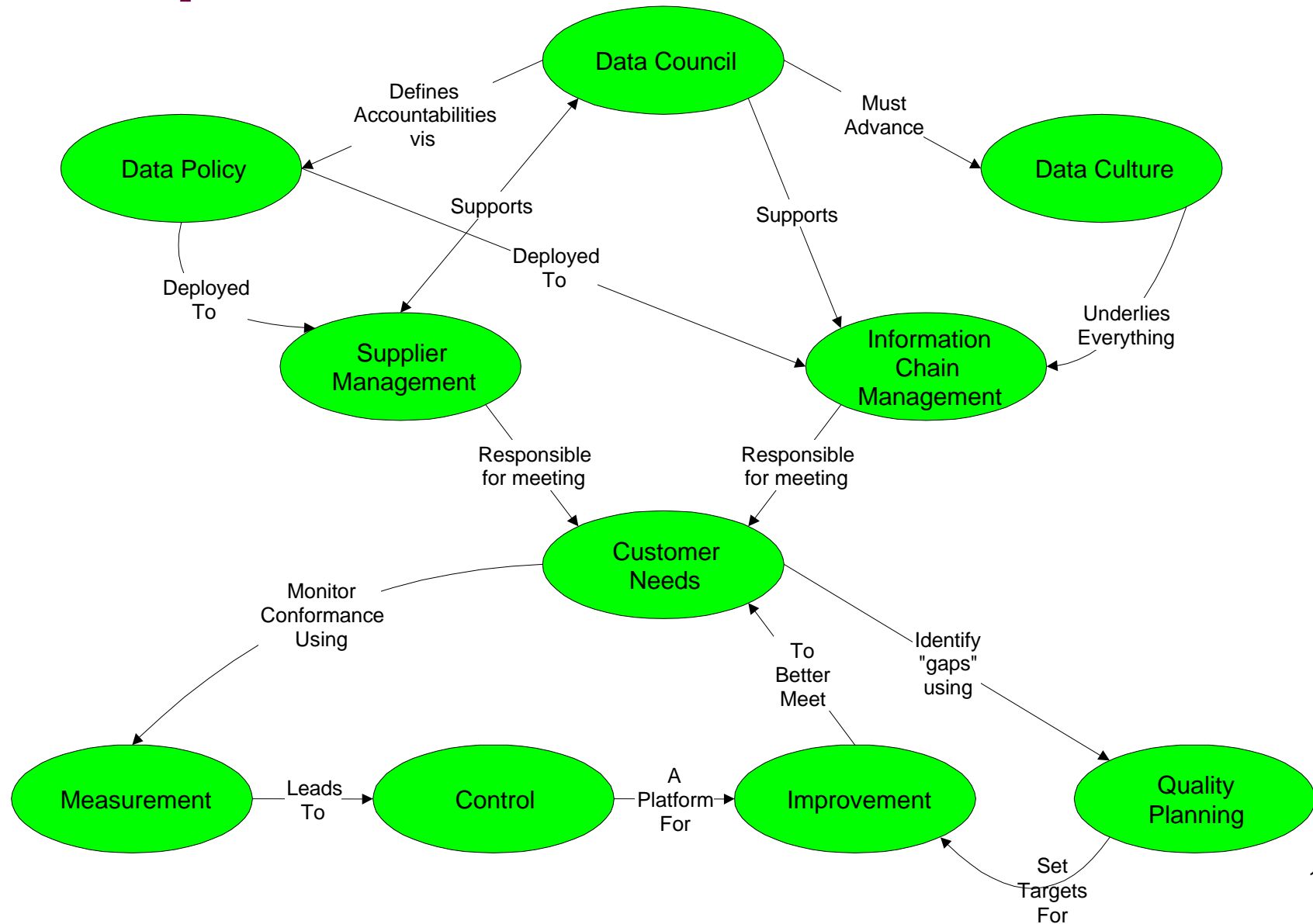


Second-Generation Data Quality Systems™

Those with the highest quality data manage the data culture. They:

- Distinguish “I” from “IT.” They recognize that automating a poorly-defined and -managed process is ill fated.
- Start small. Early wins.
- Actively manage change.
- Avoid unwinnable battles, especially early on.
- Recognize data as business assets.
- Build data quality in:
 - To the organization
 - To new systems
 - To people’s psyche

Components reinforce each other





Leadership and management are the heart and soul of data quality

So on the next several slides, we describe a bit more about the needed elements:

- Data Council
- Data Policy
- Data Culture
- Information Chain Management
- Data Supplier Management



Data Quality Council

Definition: The senior management body charged with executing the data quality policy at the highest level. Responsible for setting quality goals, selecting “projects,” providing training, and so forth. Responsible for estimating costs and benefits of quality function. May also feature a hierarchy of functional Councils.

Motivation/Advantages:

- Emphasizes senior management commitment to quality.
- Provides the cross-functional coordination and support necessary to carry out the policy and projects.

Second-Generation Characteristics:

- Very senior, with broad representation.



Data Quality Policy

Definition: A statement of management's intent regarding data and information quality,* the organization's long-term data and information quality improvement objectives, and specific management accountabilities for pursuing the intent and achieving the objectives. The policy is intended as a "guide for managerial action."

Motivation/Advantages:

- ❑ Forces organization to think broadly and deeply about quality.
- ❑ Provides insiders and outsiders a superior form of "predictability."
- ❑ Broad communication and alignment.
- ❑ Reduces "lone ranger" mentality.

Second-Generation Characteristics:

- ❑ Recognizes data and information "as business assets."
- ❑ Delineates accountabilities along information chains.



Data Supplier Management

Definition: The overall program for managing suppliers, including selecting suppliers, ensuring that these suppliers understand what is expected, measuring performance against these expectations, and making improvements to close gaps.

Motivation/Advantages:

- Much data comes from suppliers. It is too difficult to find and correct errors downstream.
- Predictable input into information chains.

Second-Generation Characteristics:

- “Partnerships” built with the most important suppliers.
- Expectations documented.
- Data quality measurements made by suppliers and regularly communicated to customers.
- Focus on “rate of improvement” rather than actual level.



Information Chain Management

Definition: Management infrastructure and technique intended to ensure accountability for the performance of cross-functional information chains.

Motivation/Advantages:

- Data and information cross organizational boundaries as Information Products are created.
- Most “problems” and/or opportunities occur on boundaries.
- Proven methods for making and sustaining improvements.
- “Control” yields predictable performance.

Second-Generation Characteristics:

- Information chains broadly defined.
- “Owners” have wide latitude for action, control, improvement.
- Customers and their needs documented and communicated.
- Overall measurement of most important quality dimensions.
- Continuous improvement.

Management of the Data Culture

Definition: Management infrastructure and technique intended to ensure that the enterprise is ready for and can adopt and make use of second-generation techniques, that promote management of data and information as business assets, and that obviate or mitigate power struggles.

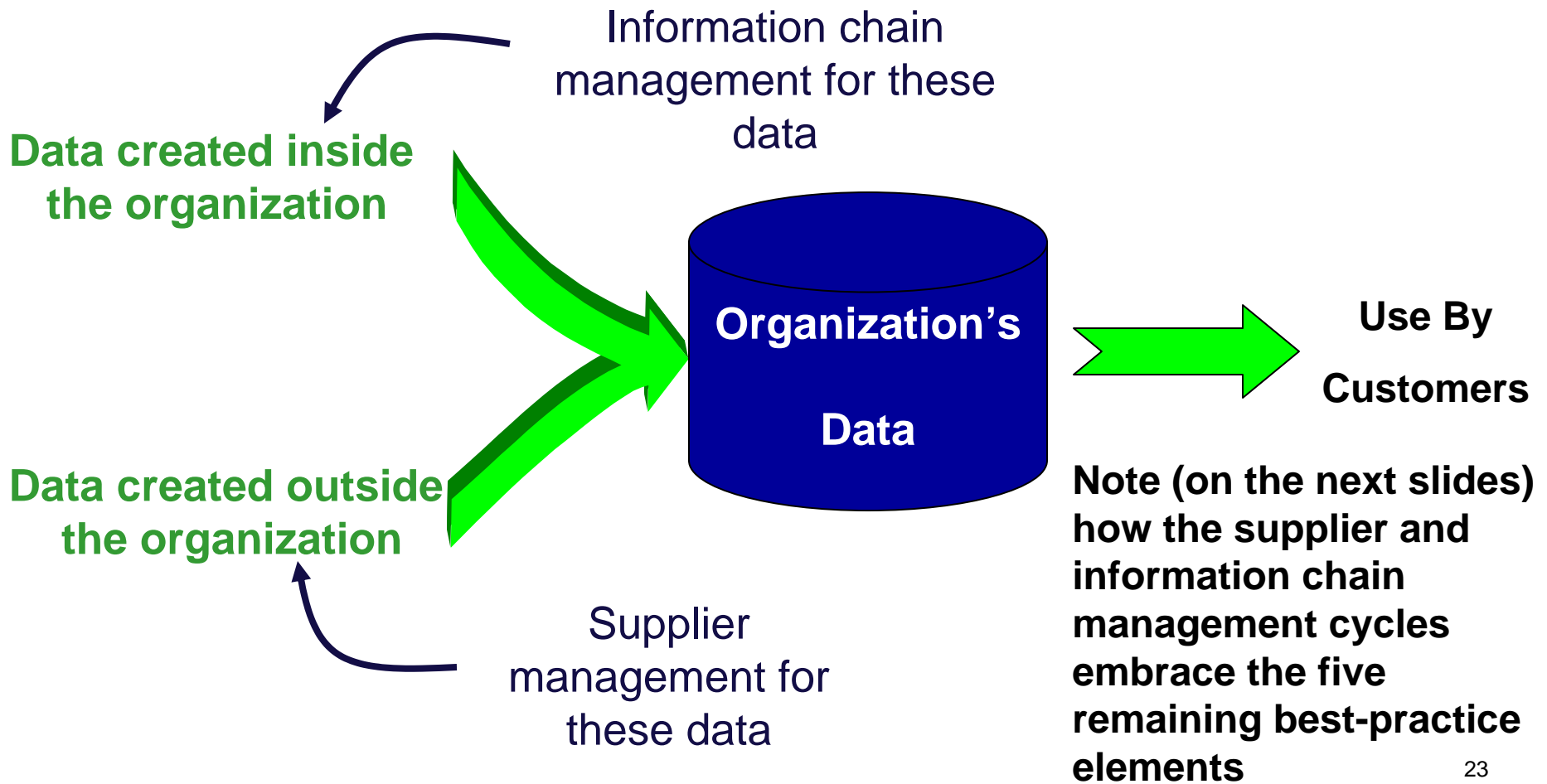
Motivation/Advantages:

- ❑ Most enterprises/organizations have first-generation data quality systems. Second-generation systems require them to think and act differently.
- ❑ Experience shows that change is always risky, but risks can be managed and/or reduced.
- ❑ Data engender politics and passions like no other resource.

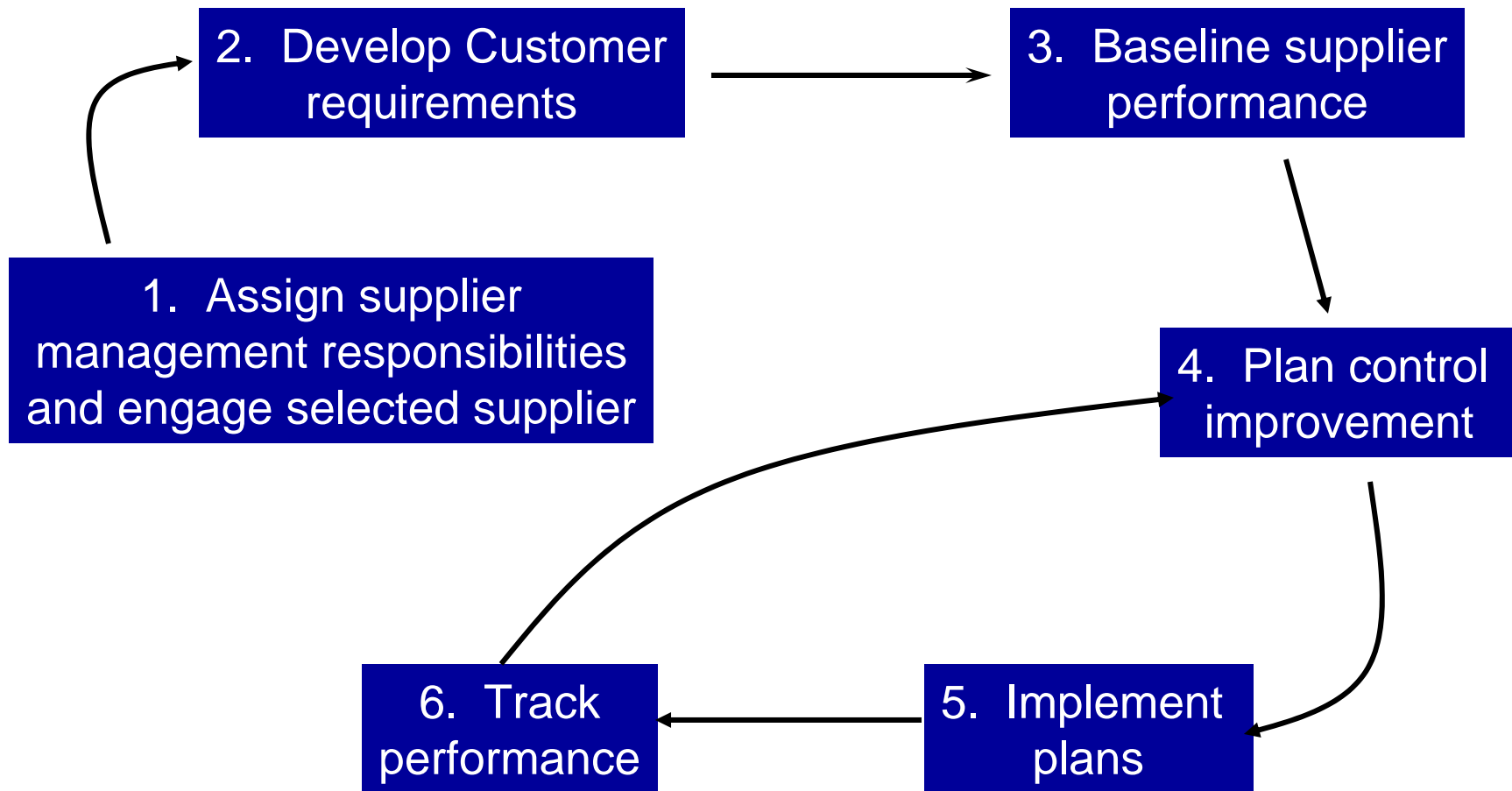
Second-Generation Characteristics:

- ❑ The responsibility of the Data Council.
- ❑ Change actively managed and based on an accepted change model.
- ❑ Leaders actively supported. People given adequate support as their jobs change.
- ❑ Unwinnable battles avoided.

Apply information chain management to data created inside and data supplier management to data created outside the organization

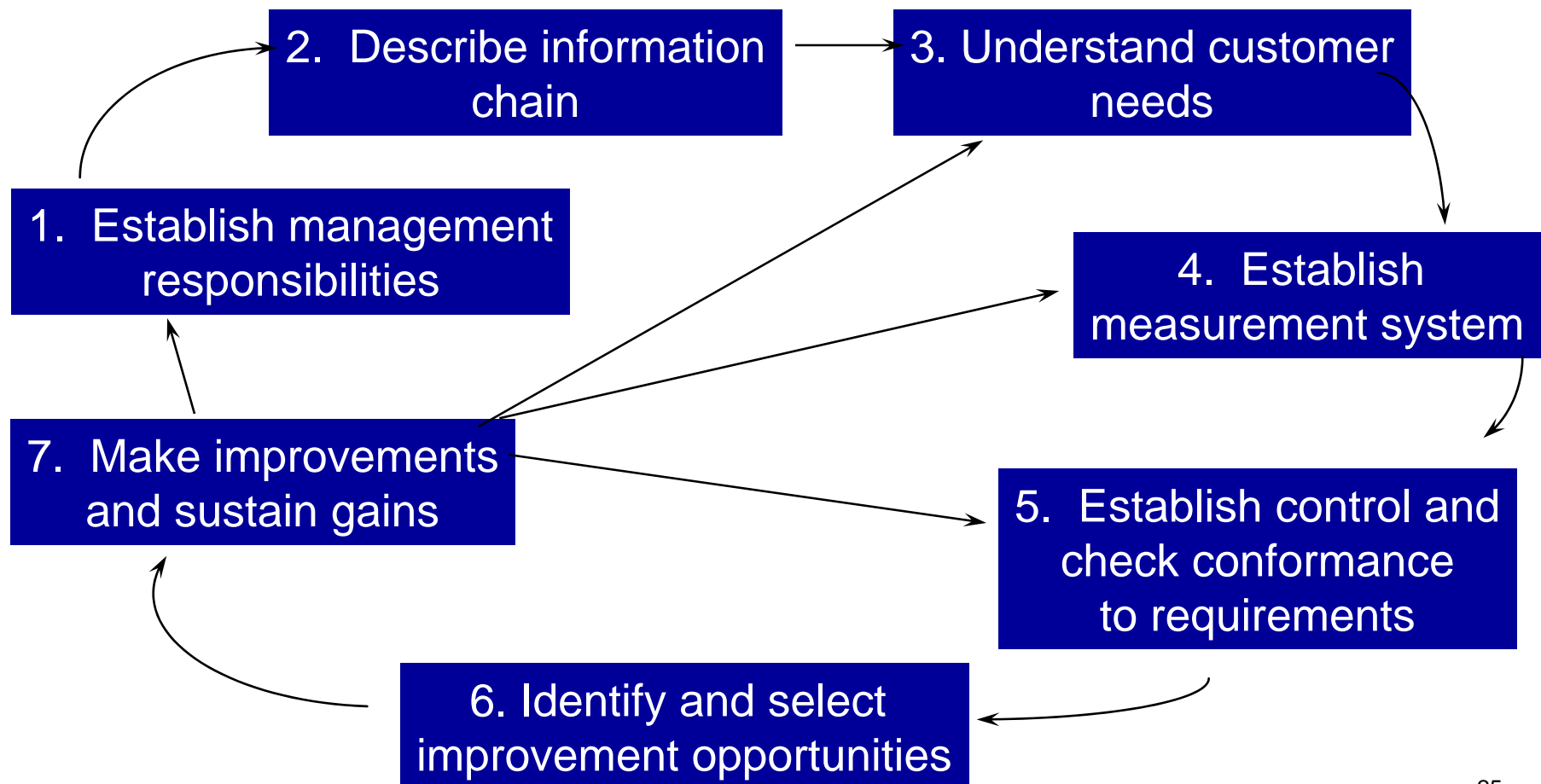


Data Supplier Management Cycle

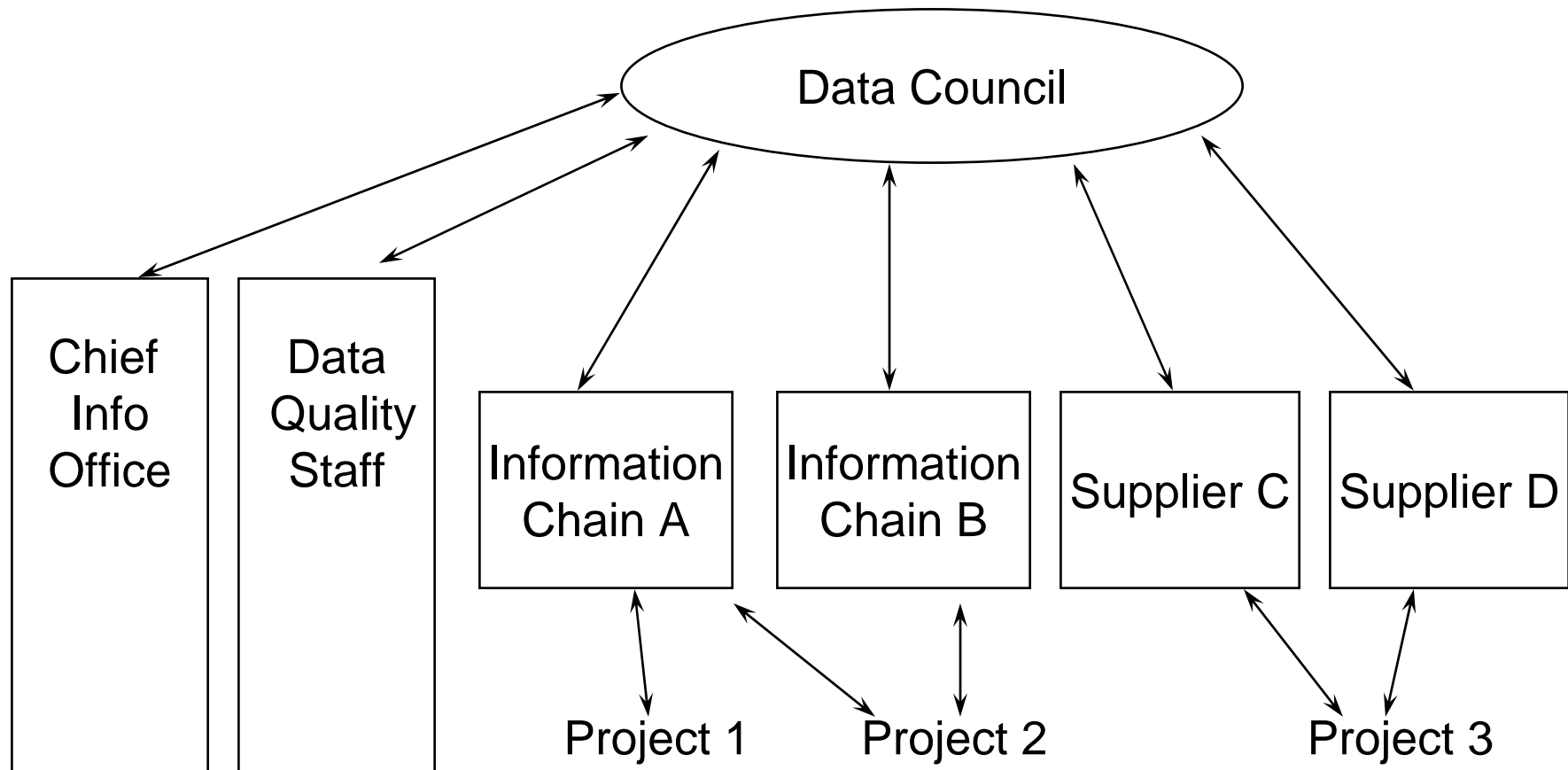


/Figures F

Information Chain Management Cycle



Proposed Governance Model for Data Quality*



*overlaid on current organization

Data Quality Planning

Definitions:

At the enterprise level: An annual process of setting quality goals or targets for quality levels and/or improvement and putting in place the means to achieve those goals.

At the “project” level: A team process that creates or replans new information products, information chains, or controls to meet specific customer needs.

Motivations/Advantages: Helps assure that information chains can consistently meet customer needs.

Second-Generation Characteristics:

- Structured methodology followed.
- Goals are the responsibility of the quality council.
- Project teams chartered by quality council.

Note: We specifically include reengineering as a quality planning technique, implemented (only) when *major* changes are required.

What Did He Say?



Any other questions?