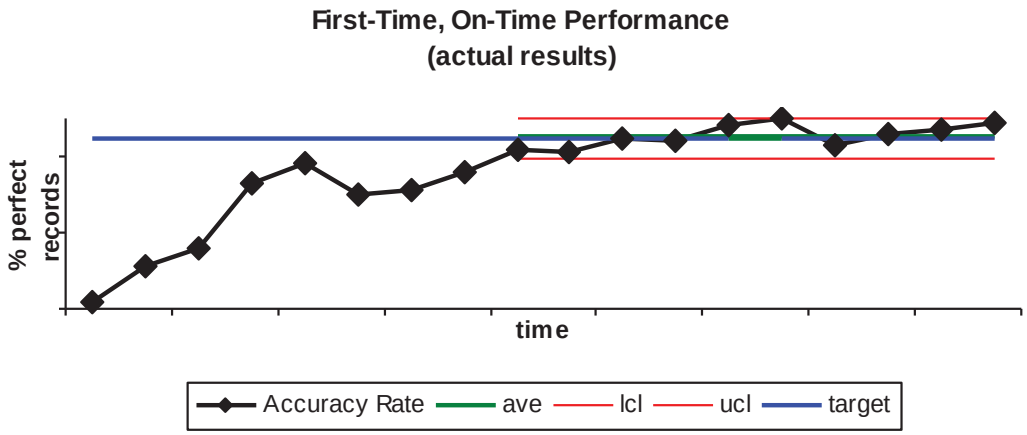
A Practitioner's View of the Really Big Data Quality (Research) Issues

Thomas C. Redman, Ph.D.
the Data Doc
Navesink Consulting Group
At ICIQ-2011,
University of South Australia
November 19, 2011
www.dataqualitysolutions.com

Agenda

- Metaphorically, (IMHO) most organizations are in the Stone Age when it comes to data (quality management).
- The late Stone Age is achievable! And the benefits are enormous!
- The failure rate is high. Why?
  - There are some real gaps in our thinking about "organization"
- (IMHO) the keys to getting to the next level lie in exploiting the notion "that data are assets."
  - Business models for putting data to work.
  - Leveraging, and accommodating properties of data unlike other assets.

Those who apply diligent efforts (almost) always improve data quality. And benefit!



Each error not made saves an average of $500.

This amounts to millions quickly!

So, why doesn't everyone have good data?

- Lazy
- Don't believe the business case
- Social, political, organizational

Big DQ Research Issue: The Business Case for DQ

Current state:

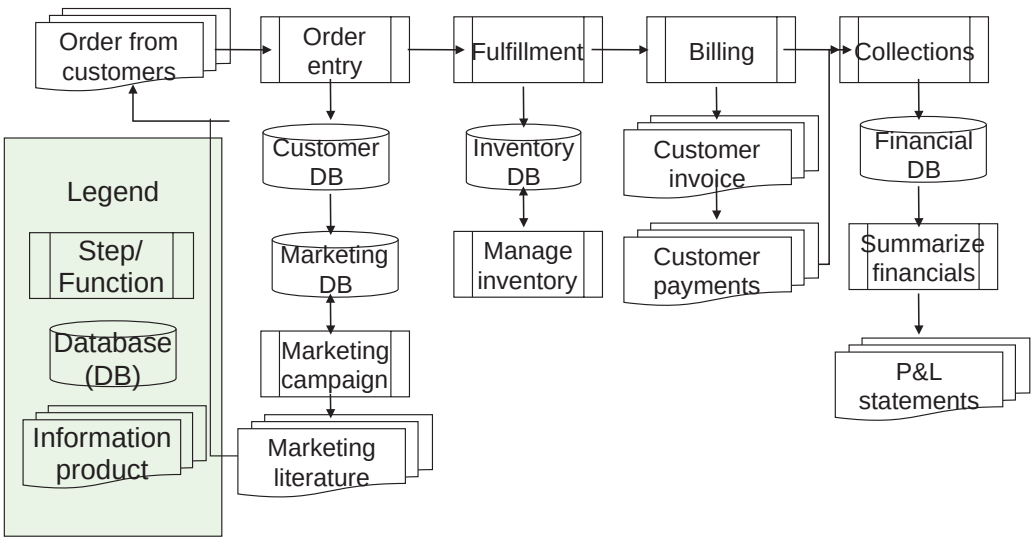
- By and large, our business cases are unconvincing.
- Decent estimates of cost only
- Strongly suspect costs are of minor consequence

Big DQ Research Issues- Business Case:

- Better ways to estimate costs
- Good ways to estimate "lost opportunity"

Figure 2.1, Redman, *Data Driven: etc*

"Management" and data flow are misaligned



Big DQ Research Issues: Organization– 1

1. **Size of effort:** How many people are needed?
2. **Organization Structure:** Where do they work?
   - What is the best (or even a better) overall organizational structure for data?
   - Does data demand entirely different management structures (i.e., process orientation)?
   - To whom does "data" report?
3. **Decision Rights:** What do they do?
   - What are the (high-level) job descriptions?
   - And decision rights?

Further on size of effort:

Consider management of people:

- An "enterprise" HR function with a very senior head
- Dedicated staff (say 1-2% in enterprise and departmental HR jobs)
- Some "line responsibilities," such as succession planning. Also, ensuring the organization has the right kinds of people skills.
- Set policy/Administer processes (such as performance review

BUT line managers do most of the actual human resource management!!

Size of effort, cont.

Contrast with the typical data organization

- Often "buried somewhere" (too often IT)
- …with too few people
- ….with technical skills (only)
- …..focused on short-term problems
- ……without adequate political capital, process, and/or support.
- ……..etc, etc, etc!!

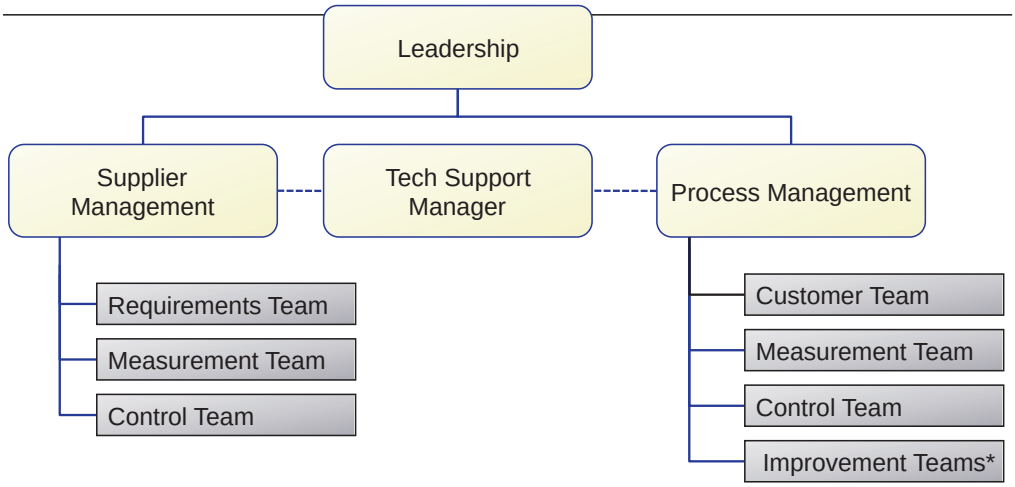
Big DQ Research Issues: Organization– 2

4. **Management Accountabilities:** Since everyone "touches" data in some way and so can affect their quality, how should the management accountabilities be defined and deployed throughout?
5. **Connecting Data Customers and Suppliers:** How should an organization ensure that data creators understand customers' quality requirements?
6. **Interaction with other management systems:**
   - Are there special considerations for "management data" (used to run the organization)?
   - How does the management system for data interact with the other management systems?
   - Ex: HR How does an organization build data into the human resources management system?

Foundations on which to build

- Federated models for managing HR and capital.
- "Current best" organizational structure
- Organizing for management data throughout history
- Cybernetics applied to management (Stafford Beer)
- Individual data processing abilities of individuals (Elliot Jacques)
- Social networks

Fundamental Organizational Unit for Data Quality



*QI facilitator is a permanent role, supporting a series of improvement project teams, which disband when their projects complete

From Stafford Beer*

Separate "management systems" needed to:

1. Manage (individual) operations
2. Ensure balance among departments
3. Create synergy
4. Innovate
5. Provide organizational identity

**The Heart of Enterprise*, Wiley, Chichester, 1979.

To the next level

Q: When did quality become important in manufacturing sectors?

A: When markets demanded it.

What Does "Manage Data Assets" Really Mean?

Generally recognized as business assets:

- Capital, in its various forms
- People, including the knowledge in their heads.

Organizations naturally manage their assets…

- They take care of them (e.g. quality)
- They put them to work, to make money.
- They evolve their management systems to account for the special properties of each asset.

A Note on Market Demands

- People and organizations have always wanted "more and better" data.
- Historically, the elite took steps to hoard data.
- Since the rise of democracy, some of their grip has been broken.
- Sheer demand continues to grow and is in little doubt:

"Inside IBM, we talk about 10 times more connected people, 100 times more network speed, 1,000 times more devices, and a million times more data."*

*Lou Gerstner, quoted in McDougall, P., "More Work Ahead," *Information Week*, December 18-25, 2000, p. 22.

A Note on Market Demands-2

- To borrow from Twain,
  "the difference between the right data and the almost-right data is like the difference between lightning and a lightning bug."
- People and organizations expect:
  "exactly the right data and information in exactly the right place at exactly the right time and in the right format to complete an operation, serve a customer, make a decision, or set and execute strategy."

Informationalization

Basic Idea: Enhance existing products and services by building data and information in.

Customer Needs Fulfilled: Simplicity, integration, etc.

Industrial Age example:

Information Age examples:

- Auto makers are now including GPS Navigation systems.
- NC State is re-designing the hospital gown, building a thermometer and other sensors in.

Infomediation

Basic Idea: Help people find the data and information they need

Customer Need Fulfilled: Waste less time.

Industrial Age example: Travel agents

Information Age example: Google

Big Data Quality Research Issues: Monetizing Data

So far, I've identified fifteen ways to put data to work

Provide Content

- New Content
- Re-package
- Informationalization
- Unbundling
- Exploiting Asymmetries
- Closing Asymmetries

Facilitators

- Own the Identifiers
- Infomediation
- Data mining/ Analytics/Big Data
- Privacy and security
- Training
- New Marketplaces
- Infrastructure technologies
- Information appliances
- Tools

Big DQ Research Issues:

1. Working out the business models for each: the "heavy lifting" of the Information Age!
2. New ways to put data to work

Big DQ Research Issues: Six-sigma capabilities for DQ Measurement and Control

Sooner or later, markets demand extremely high-quality:

- Can't support measurement more precise that +/- .001
- Measurements:
  - Eyeballs, comparison to reality: Time-consuming and expensive.
  - Based on "counts" of failed business rules
- Controls largely based on business rules
- Don't even know how to rate adequacy of business rules

Big DQ Research Issue: Measurement and controls adequate to support 6-sigma levels of quality (3.4 ppm)

- In the face of "intangibility"

Direction of solution: Information theory (e.g., telecommunications)

Data possess properties unlike any other asset. We're just beginning to understand these properties

Property	Example Implication for DQ
Data multiplies	Sheer rate of growth strains ability to manage them
Data are more complex than they appear	Much (data model, correct values, presentation) has to go right
Data can be digitized	They are (technically) easy to share. And steal.
Data create value "on the move"	Data sitting in a database are profoundly uninteresting
Data are intangible	They have no physical properties, complicating measurement
Data are organic	Data morph as they move to suit different needs
Data are the means by which organizations encode knowledge	They have very long lifetimes.
Data are subtle and nuanced. They have become the organization's *lingua franca*	The distinction between the "right data" and "almost right data" is akin to the distinction between "lightning" and a "lightning bug" (after Twain)
Each organization's data are uniquely its own	They are the "ultimate proprietary technology"

Power / Data Sharing / Ownership –1

In the Information Age, possession of data conveys power.

Power / Data Sharing / Ownership - 2

Though universally praised, data sharing is an exception.

- Recommended book:
  The 48 Laws of Power (Greene and Elffers, Viking, 1998)
- Many of the laws seem to argue against sharing data!

Big Data Research Issue: Data sharing

Current State:

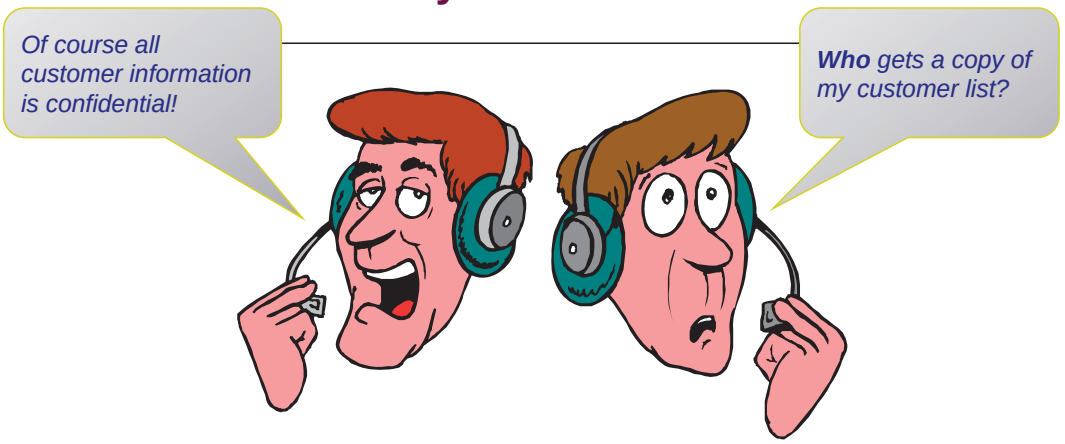
- Unlike other assets, data can be shared. The potential benefits are enormous!
- But sharing is counter to human nature (at least for the powerful).
- In organizations, we ask people to "own" problems/opportunities. But "ownership" of data conveys rights that are counter to sharing.

Big DQ Research Issue: How do we sort all this out?

Possible directions:

- Incentives?
- Internal data markets?
- New organizational forms? (a close reading of Stafford Beers suggests a separate management system is needed)

Privacy is a Wild Card



Accepted practices, legal frameworks and traditions regarding privacy have not been developed.

Prescient Quote

"Privacy will be to the Information Age what product liability was to the Industrial Age."*

In the United States anyway, consumer protections continue to grow even today.

Common Data Definitions

Common definitions of data have proven remarkably difficult.

Data are subtle and nuanced and have become the organization's *lingua franca*

Never underestimate the importance of local knowledge.

HSBC
The world's local bank

An Organization's Data are Uniquely its Own

- No other organization has, or can have, the same data.
  - Created within.
  - Most data are simply "not for sale."
- Data are subtle and nuanced.
  - Model "customer" in unique ways that best suit it.
  - Capture and utilize unique "facts."
  - Processes to capture unique data are also difficult to copy.
- Eventually, of course, some data do become standardized.
- Data offer opportunity for sustained advantage—and everyone knows it!

Implications

- Must be very careful about what data we standardize. Standard data has lower marketplace value.
- Should strive for greater uniqueness, novelty, and depth in data put to work in the marketplace.
- At the same time, standard data promotes both internal and external communication.
- The virtues of sharing complicate this.
- The demands of privacy complicate this.

Big DQ Research Issues: Properties of Data - 1

1. **Data Sharing**: How should an organization promote and/or demand data sharing? Under what terms and conditions? Is a special organizational unit required?
2. **Special provisions for unique data:** How should an organization identify, acquire, nurture, protect, and utilize (monetize) data that are uniquely its own?
3. **Standard Data:** How should an organization resolve the need for "standard data" to promote cross-departmental communication with the need of individuals for highly-nuanced data to complete their tasks?
4. **Meta-data:** Are special methods for development and promulgation and use of meta-data needed? If so, what are they?

Big DQ Research Issues: Properties of Data - 2

5. **Privacy**: How should the organization think through its privacy obligations and/or what is smart business?
6. **Implications for business models:** How do these properties impact the business models?
7. **Implications for organization:** Do these properties and the new business models demand fundamentally different organization structures?

Big DQ Research Issues: What am I missing?

Background:

- I can't "see around corners.
- I have no crystal ball.

Biggest DQ Research Item: What am I missing?

Final Remarks

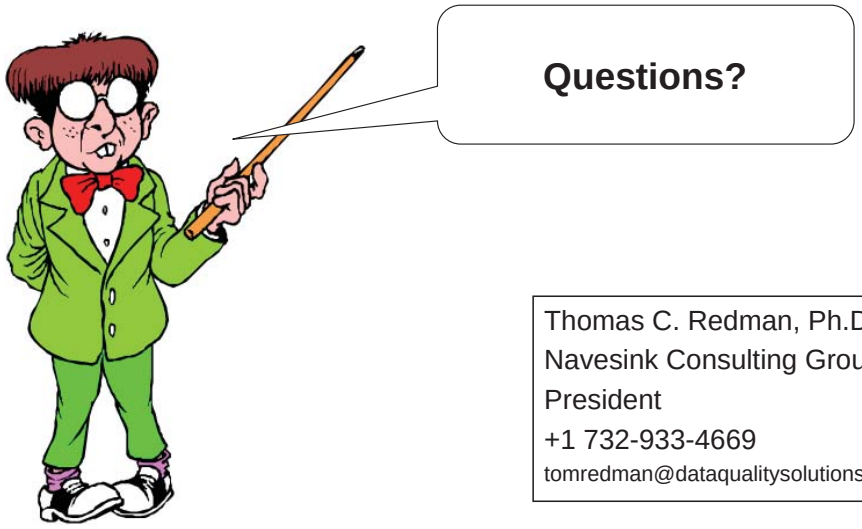
To Advance Now:

- Better Business Cases
- Fundamental organizational issues

To Get to the Next Level:

- Business Models (to put data to work)
- To account for (leverage and accommodate) the unique properties of data
- What am I missing?

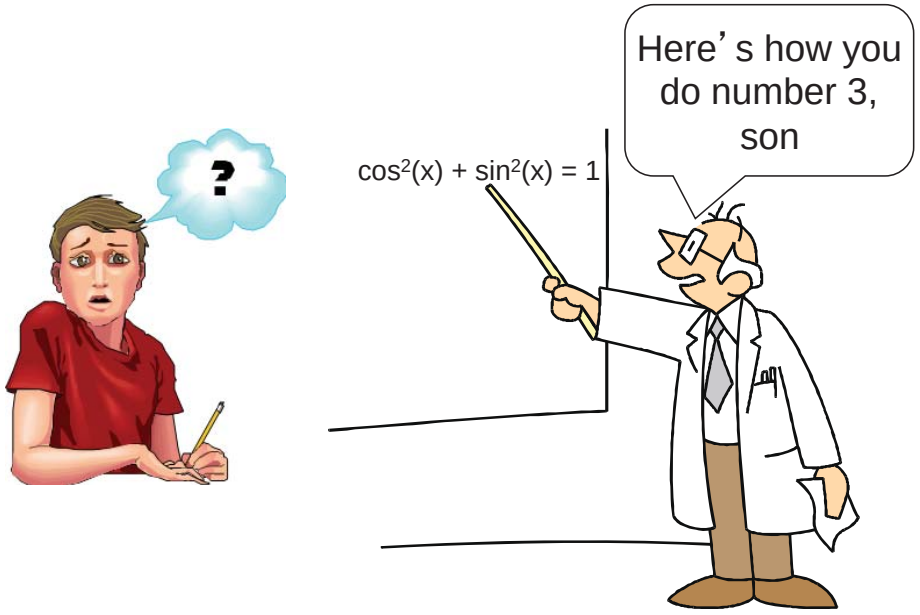
What Did He Say?



Thomas C. Redman, Ph.D.
Navesink Consulting Group
President
+1 732-933-4669
tomredman@dataqualitysolutions.com

A Database is Like a Lake

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



Ten Habits of those with the best data*

8. Formalize management accountabilities for data

9. Broad, informed, demanding leadership

10. Advance a culture that values data and data quality

7. Set and meet aggressive targets for Improvement: top-to-bottom

Senior Leadership:

Middle Management (Command):

2. Manage processes that create data (so they do so correctly)

Work is highly interconnected

3. Manage "suppliers" (both inside the Army and out) of data

Everyone who touches data = Four Basic "Tasks"

1. Focus on the most important needs (of customers)

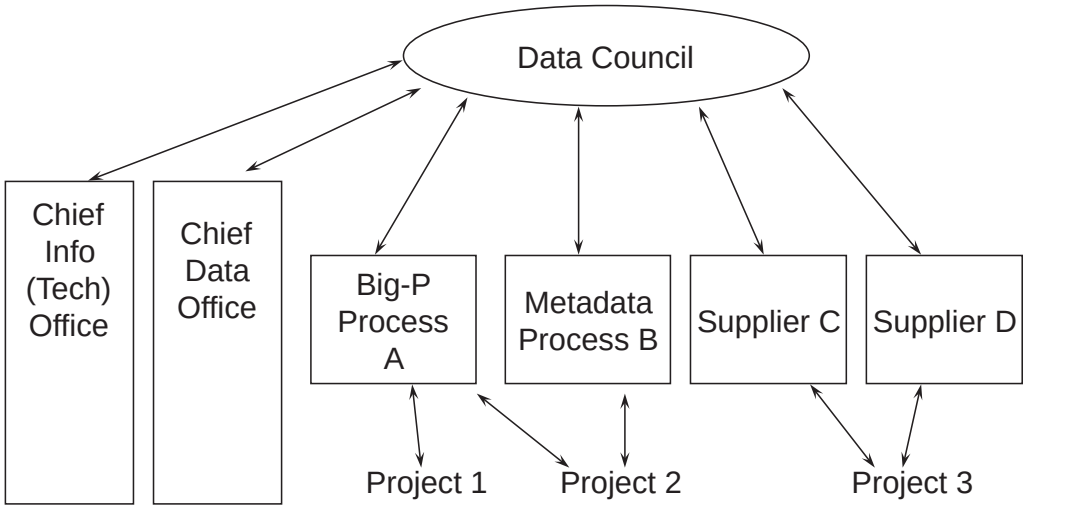
4. Measure quality levels against customer needs

5. Deploy controls, at all levels, to remain error-free*

6. Improvement: Find and eliminate root causes of error

*Ten habits of those with the best data from Redman, *Data Driven: Profiting from Your Most Important Business Asset*, Harvard Business Press, 2008.

"Current Best" Organization Structure for Data*



*overlaid on current organization

For data, "taking care" is mostly about quality

Prescription 1: Take steps to ensure that

- Possess and acquire the right kinds of data.
- People can access and understand them.
- People can trust that they are "good enough."
- They are of high enough quality to withstand market scrutiny.
- They are kept safe from loss or theft.

It is highly significant that (almost) all organizations that diligently follow many of "the ten habits" make order-of-magnitude improvements.

Putting data to work

Prescription 2: Use data to create new revenue

- Sell them directly in the market.
- Build them into other products and services.
- Use them to enhance other products/services.
- Use them to make better decisions.
- Use them to improve the day-in, day-out running of the business.

Critical point: Management must explicitly think through how they will put data to work in creating new value.

Adjusting the management system

Prescription 3: Recognize that data have unique properties

- Example: Unlike other assets, data can be shared
- Most important: Data are the only asset that are uniquely an organization's own. The "ultimate proprietary technology."

Prescription 3, cont: Evolve organizational structures, roles, and responsibilities as a result.

- Counterexample: Chief Information *Technology* Officer

Consultant's exercise: Fire!

You can save only one of the following:

- Antique French Desk.
- Brand new PC, with all the bells and whistles.
- Only copy of the organization's fifty biggest accounts.

The Data Doc's Response: Finally!

The paradox

- How do we reconcile the fact that everyone knows data are critical assets with the fact that organizations don't manage them as such?

- Hypothesized Answer: They don't know what to do.