Overcoming Big Data Quality Challenges and Emergence of the CDO

2013 MIT Research and Development Conference
November 13-14, 2013

Stuart Madnick
Sloan School of Management
MIT

Track 6: Safety, Risk, and Quality
Agenda

1. Why is Big Data a Disruptive Force?
2. Big Data Quality Challenges and the MIT Information Quality Program
3. Chief Data Officer (CDO) as an Approach to Managing and Exploiting Big Data
4. A Cubic Framework for Chief Data Officer
Part 1: Why is Big Data a Disruptive Force?
McKinsey strategic consulting firm: “Big Data is the next frontier for innovation, competition, and productivity” 1

• “The amount of data in our world has been exploding …
• … So called ‘Big Data’ will become a key basis for competition, underpinning new waves of productivity, growth, innovation, and consumer surplus…
• … Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers…
• … The rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future…”

1 http://www.mckinsey.com/mgi/publications/big_data/
It is now “official” that Big Data is important!
So ... What is Big Data: Simple View (2 V’s)

• $V_1 = \text{Volume}$
  – Not really all that new
  – I was a founding Trustee of the Very Large Data Base (VLDB) Foundation – started around 1975

• $V_2 = \text{Velocity}$
  – Somewhat new, popularized by “streaming” data
    • Such as high-speed satellite data that must be processed as it arrives
What is Big Data: Expanded View (3 V’s)

- $V_3 = \text{Variety (multiple dimensions)}$
  - Multiple sources (of similar data), but ...
    - Differences in format
    - Differences in semantics
  - New types and sources of data
    - Social media
    - Location-based data
  - Multiple sources (of different data)
    - Combining traditional data (such as prices of food) with social media data (sentiment) to forecast food riots
New Tools Beget Revolutions and Disruptions
Disruptions:
Often Require Doing Things Differently

• Often changes your standard IT activities

• Technical changes
  – New types of data
  – New ways needed to process the data

• New Business Strategies and Applications
  – To understand the data
  – To exploit the data
  – To adjust the organization appropriately

• Need to think differently both technically and strategically
New Sources of Data (some examples)

- Web traffic: Clickstream/ Page views/ Web activities
- Web links/ Blog references
- Search engines: Google/ Bing/ Yahoo
- Social media: Facebook / Twitter feeds
- Location and Activity: Mobile phone/ GPS
- Email messages
- Transactions: ERP/ CRM/ SCM
- RFID (Radio Frequency Identification), Bar Code Scanner
- Real-time: Machinery diagnostics/ engines/ equipment
- Automated scientific equipment: DNA sequencers
- Financial transactions: Stock markets / foreign exchanges
- User generated content: Wikipedia updates
- Open Linked Data
- Online repositories

A Few Example Applications
Search Engines: Ability to do Prediction

• **Insight:** *We know what you are thinking!*

• Google Search Foreshadows Housing Prices and Sales

• Data sources used in economics have substantial lag and high level of aggregation -> difficult to use for real-time predictions

• Data from search engines like Google provide highly accurate way to predict future business activities

• *Example:* Predict housing market trends
  – Produced much better results than conventional models

• This is a form of *implicit* Collective Intelligence

• Also used to identify emerging “hot” research – before known (from analyzing 100,000’s of research reports)
Using Detailed Sensor & Social Data

• **Insight:** *We may be able to know things about you that even you don’t (yet) know …*
  – *Based on work by Sandy Pentland and MIT Media Lab*

• **Many “sensors” on smart phone:**
  – Know who you call
  – Know where you are
  – Know who is next to you

• **Using accelerometer**
  – Know your movements -> Many applications
  – Interesting example: Anticipate medical developments, e.g., depression, post-traumatic stress disorder, and other mental illnesses
Social Media Predictive Analysis: 2011 London Riots

Insight: We know what you are thinking about doing!

- Based on work by Stuart Madnick research group
  “Bottom up” combined with “Top down” analysis

Notion of “Narratives”

MPS Incident Data

Social Media Data and Analysis Tool Provided by Crimson Hexagon
Comparison with Data

Oppression Narrative Building

Protester Incidents Aug 6

Narrative of Oppression

Simulated Open Incidents
Some other potentially controversial uses

- **Insight:** There are many things that can be learned about you by studying your social network ...
  - Raises lots of privacy-related issues ...

---

**Project ‘Gaydar’**
At MIT, an experiment identifies which students are gay, raising new questions about online privacy

By Carolyn Y. Johnson
Globe Staff / September 20, 2009

It started as a simple term project for an MIT class on ethics and law on the electronic frontier. (Full article: 1768 words)
Social Media Data -> issues and opportunities...

“Personal data is the new oil of the Internet and the new currency of the digital world”,
Meglena Kuneva, European Consumer Commissioner
Revisit: What is Big Data: Expanded View (4 V’s)

- \( V_4 = \text{Value} \)
- Having Lots of Data is not very interesting, unless Value can be obtained from the data.
- We have already seen some examples of value:
  - Predict housing prices
  - Discover early stages of certain diseases
  - Anticipate riots and social disruptions
  - Etc. (more later)

- How does an organization get value from Big Data?
  - Major focus of the rest of our presentation
  - But first a brief aside ...
What is Big Data: Expanded View (5 V’s)

$V_5 = \text{Veracity! (truthfulness)}$

Naïve view

- Since there's so much data, don't need to worry about quality -- 'a few pieces of bad data' can be ignored
- *Au contraire*: because the sources are usually so new and so large, we often know very little about the data
Part 2: Big Data Quality Challenges & MIT Information Quality Program
An Example of Misunderstanding Large Quantities of Data

• During a previous housing market crisis
  – *Boston Globe*\(^1\) reported that there was a **significant increase in house sales** – Sounds like Good news!
  – But, upon later investigation, it was learned:
    • Basis was number of changes in ownership filed with Registry of Deeds
    • Normally that would indicate that house was sold to someone
    • But, changes also occurred when bank foreclosed on a home!
    • **What really happened was the banks were foreclosing at an increased rate**
  
... not that more people were buying homes.

\(^1\) The *Boston Globe* is the most authoritative source of information in the world.
The Importance of the “Veracity” Aspect: Evolution of Data & Information Quality

**Journals** *
- 2007 ACM Journal on Data and Information Quality (JDIQ)

**Conferences and Certification Programs** *
- 1996 International Conference on Information Quality (ICIQ)
- 2002 MIT-IQ program for Executives
- 2003 IQ-1: Principles and Foundations
- 2007 IQ Industry Symposium
- 2010 QIBRAS Conference, etc
- **2011 First CDO Forum** . . . .

**Books** *
- *Data Quality* (2000)

**Articles** *
- 1990 Polygen Data Quality Model *(VLDB + ICIS)*
- 1996 Beyond Accuracy
- 1998 Managing Information as a Product, etc …

**Research Projects** *
- 1988 Total Data Quality Management Program (TDQM)
- 2002 MIT Information Quality (MITIQ) Program

**Education**
- MS IQ and IQ PhD
- Degree Programs

* Not complete list
Some Data Quality Research Areas

- Data Quality is multi-dimensional
- Organizational Data Quality assessment
- Interplay of Data Quality and Data Semantics

- Manage information as a product
- Data integrity analysis
- Data Quality root cause analysis
- Data Source/Provenance – mathematics of DQ
What is Data Quality?

• Naïve / Conventional view:
  
  Data Quality = Accuracy

• Research finding:
  
  Data Quality *Goes Beyond* Accuracy

  Initial survey of data users resulted in over 100 different data quality dimensions!

• What are some other dimensions?
**Data Quality Dimensions:**

16 Key dimensions, organized into 4 categories

<table>
<thead>
<tr>
<th>DQ Category</th>
<th>DQ Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic DQ</td>
<td>Accuracy, Objectivity, Believability, Reputation</td>
</tr>
<tr>
<td>Accessibility DQ</td>
<td>Access, Security</td>
</tr>
<tr>
<td>Contextual DQ</td>
<td>Relevancy, Value-Added, Timeliness, Completeness, Amount of data, Ease of manipulation</td>
</tr>
<tr>
<td>Representational DQ</td>
<td>Interpretability, Ease of understanding, Concise representation, Consistent representation</td>
</tr>
</tbody>
</table>
Organizational DQ Assessment

Many different roles involved with data in an organization ...

- Method: Questionnaire to Assess Perceptions of Data Quality
- Analysis: Statistical Significance, Statistical Reliability and Statistical Validity (Convergent Validity and Discriminant Validity)
Organizational DQ Assessment: Some sample results

![Graph showing IQ Assessment and Importance](image-url)
Organizational DQ Analysis: Some sample results

Free of error: Agree

Ease of Manipulation: Disagree
Interplay of Data Quality and Data Semantics

**Daimler Benz (DCX) Financial Data**

<table>
<thead>
<tr>
<th>Source</th>
<th>P/E Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>11.6</td>
</tr>
<tr>
<td>Bloomberg</td>
<td>5.57</td>
</tr>
<tr>
<td>DBC</td>
<td>19.19</td>
</tr>
<tr>
<td>MarketGuide</td>
<td>7.46</td>
</tr>
</tbody>
</table>

Which one is correct? Why?
Connection between Big Data & Data Quality

• **Remark from many executives:**
  “I now have more and more information, that I know less and less about...”

• **Big Data provides:**
  – Even more data
  – From even more diverse sources

• **To get true and effective value from Big Data**
  – It must be **high quality** Big Data
In 1805, the Austrian and Russian Emperors agreed to join forces against Napoleon. The Russians promised that their forces would be in the field in Bavaria by **Oct. 20**.

The Austrian staff planned its campaign based on that date in the **Gregorian calendar**. Russia, however, still used the ancient **Julian calendar**, which lagged 10 days behind.

The calendar difference allowed Napoleon to surround Austrian General Mack's army at Ulm and force its surrender on Oct. 21, well before the Russian forces could reach him, ultimately setting the stage for Austerlitz.


**Bottom Line:** Increasing opportunity to gather and integrate information from many diverse BIG DATA sources ... But both a technological and Strategy change.

• But ... need an executive (a CDO) to be sure it is understood and used effectively.
Part 3: Chief Data Officer (CDO) as an Approach to Managing and Exploiting Big Data
New CDOs are emerging to Manage and Exploit Big Data

- CIO leads business innovations in the world of big data
  - Profit from 5V’s: Volume, Velocity, Variety, Value, Veracity
  - Avoid creating many Big Data silos and duplication of effort
  - Exploit data for external, strategic opportunities

- CDO is increasingly emerging to be the focal point in data practice (not systems)
What is the CDO?

- We broadly define the Chief Data Officer as:
  Those who have the CDO title and those who are performing the CDO function

- Wikipedia describes the CDO as:
  A corporate officer responsible for enterprise-wide governance and utilization of information as an asset
Who is Appointing CDOs?

- **Government**: US Army, UK Royal Army, Federal Communication Commission (FCC), Veteran Affairs (VA), Federal Reserve Board (FRB), FRB of New York, UK Government, Baltimore City Mayor’s Office

- **Financial services**: Citigroup, Bank of America, Morgan Stanley, VISA International, Conning, TD Ameritrade, Boa Vista – Brazil, GBM – Mexico, HSBC, Conning, Standard Bank Group

- **Health Care**: Seattle’s Children’s Hospital, National Institute of Health (NIH)

- **IT companies**: Microsoft, Yahoo!

- **Retail**: Alibaba Group – China, eBay, Amazon

- **Customer Science**: Dunnhumby
To whom do CDOs report?

Our research and practice show that CDOs have various reporting lines:

- **Tier 1:** CDO reports to the CEO/COO directly (FRB, Seattle’s Children’s Hospital, Boa Vista-Brazil)

- **Tier 2:** CDO reports to CIO, CMO, CFO, CRO, etc. (US Army, Microsoft, GBM-Mexico, Conning)

- **Tier 3:** CDO is not in the C-Suite, but has the highest level job to organize and manage data (TD Ameritrade, FCC)
Example: FRB - CDO Office Strategic Importance

✓ Data and data management play a critical role in fulfilling the Federal Reserve Board’s (FRB) mission

✓ Stated FRB Strategic Objectives:

1. Improve data governance by establishing a new Office of the Chief Data Officer (CDO) and ensuring clear roles and responsibilities among the Chief Data Officer, the Board Data Council, and data users

2. Ensure that all enterprise data are handled, processed, stored, and disseminated by professional data management groups

3. Strengthen the Board’s data environment by establishing an infrastructure to share data and improve opportunities for data integration that supports the Board’s research and analytical capabilities

Part 4:
A Cubic Framework
for the Chief Data Officer

Building Organizational Capability
Current Practice:
What are organizations doing to exploit big data?

• Over 500 leading global executives reveal that most companies are **beginning to learn** about managing big data.

• The companies who have **an executive** responsible for data management achieve **higher financial performance** than their peers.

*Source: Industry survey by the Economist, 2011*
The CDO Leads Building
Data-Grounded Organizational Capability

Three Dimensions of CDO Engagement

• **Collaboration Direction** (Inward vs. Outward)
• **Data Space** (Traditional vs. Big Data)
• **Value Impact** (Service vs. Strategy)
1. Collaboration direction: Direction of the CDO’s engagement

- **Inward**: Towards internal business processes associated with data flow
  - Often closely associated with collaboration within internal business divisions or functional areas
  - Example initiatives: Data quality, data governance, architecture, analytics

- **Outward**: Towards external value chain and environment
  - E.g., customers, partners, government agencies, markets and environments.
  - External reporting, data sharing, data standards/integration consortium
2. Data Space: Traditional vs. Big Data

- **Traditional Data**: Transactional data that is typically in relational databases
  - Example: Customer information, sales data, etc.

- **Big Data**: Newer and more diverse data
  - Example: Social media data, machine-sending data, non-structured feedback data, etc.
3. Value Impact: Service vs. Strategy

• **Service**: Data delivery service
  – Example: focus on improving business operation

• **Strategy**: Exploring new strategic opportunities for an organization
  – Example: creating new markets, products, or new ways or models of conducting business
Eight Functional Roles of the CDO

Based on 3 dimensions of CDO’s engagement, the CDO take on different roles...

1. **Coordinator** (Inward, Traditional, Service)
2. **Reporter** (Outward, Traditional, Service)
3. **Architect** (Inward, Traditional, Strategy)
4. **Ambassador** (Outward, Traditional, Strategy)
5. **Analyst** (Inward, Big, Service)
6. **Marketer** (Outward, Big, Service)
7. **Developer** (Inward, Big, Strategy)
8. **Experimenter** (Outward, Big, Strategy)
CDO Roles are not Static over Time: Example of 10-year Evolvement of a CDO

Sequence of Roles:
1 Coordinator (#2)
2 Reporter (#1)
3 Architect (#3)
4 Ambassador (#4)
5 Analyst
6 Marketer
7 Developer (#5)
8 Experimenter
Overall Insights and Takeaways

• There are exciting new opportunities made possible by exploiting big data
  – Need advanced techniques and perspectives
  – Need to understand new business opportunities

• Need to ensure the quality of the information

• Need a new breed of leaders to drive big data capability for Business Process Transformation and Business Strategy Development

• A Chief Data Officer (CDO) can be critical to success

• The CDO Cubic Framework provides a systematic way to understand CDO roles and the needs of your organization

Bottom line: A new breed of executives (CDOs) can be groomed defined for effective exploitation and quality of big data to benefit the organization
Backup Slides
To be an effective **reporter**, need to be an effective **coordinator**

---

**#2: Reporter**  
(Outward, Traditional, Service)  

**Timely reporting**

Q: Is our business reporting timely and efficient?  
- Develop efficient process for external reporting  
- Implement Data warehouse for efficient retrieval and use of data for reporting and management.

“Burn that data warehouse down!”

CEO: “I need one number, not two answers to the same question”

---

**#1: Coordinator**  
(Inward, Traditional, Service)  

**Coordinated internal process**

Q: Does our internal process generate quality data for business operation?  
- Assess data quality  
- Implement dashboard to visualize repeated data quality assessment  
- Analyze root-causes of the poor quality data (information product mapping, etc.)  
- Establish data governance mechanisms: stewardship, council

---

- Improve internal business process  
- Make data quality problem visible, solution mechanism approachable to all
From coordinator to architect...

#1: Coordinator
(Inward, Traditional, Service)

- **Behavior incentive system**
- **Architecture alignment**: Align data improvement with business processes

#2: Reporter
(Outward, Traditional, Service)

#3: Architect
(Inward, Traditional, **Strategy**)

**Sustainable strategy**

Q: Do the individual members contribute to improving data?

- Reward good behavior; Develop data charter: individual data role and responsibility tied to performance evaluation
- Establish **conflict resolution council**: data policy, DQ workgroups, governance councils
- Utilize enterprise **architecture** to align business process with data improvement initiatives
From architect to ambassador...

#1: Coordinator
(Inward, Traditional, Service)

#2: Reporter
(Outward, Traditional, Service)

#3: Architect
(Inward, Traditional, Strategy)

#4: Ambassador
(Outward, Traditional, Strategy)

Interoperability & Comparability

Q: Can we share and compare data with others?

• Reach out external organizations to improve industry data exchange, sharing and comparability

• Form Industry consortium to reach out and discuss with other CDOs

• Develop data standards and business process standards

Engage industry form for improved comparability of shared data
From ambassador to developer...

#1: Coordinator
(Inward, Traditional, Service)

Q: Are we getting a complete picture?
- Learn from analysis of patient feedback data (big data)
- Develop strategic use for combining patient feedback data with hospital assessment data (e.g., HCAPS)

#2: Reporter
(Outward, Traditional, Service)

#3: Architect
(Inward, Traditional, Strategy)

#4: Ambassador
(Outward, Traditional, Strategy)

#7: Developer
(Inward, Big Data, Strategy)

Developing insights

Collaboration with external entities

Explore added value from diverse data source