Lecture#1

**Categories of Data/Information Quality (slide#18)**

Four target categories characterize and capture dimensions for data quality (Wang and Strong, 1996). The categories address

1. Accuracy of data
2. Relevancy of data
3. Representation of data
4. Accessibility of data.

Often, there are trade-offs between accuracy and other dimensions, and other suitable elements of determining the information criteria.

Wang and Strong (2012) empirical findings show that the perceived importance of IQ measurement criteria is influenced by

1. The expectation of Information and Communication Technology (ICT),
2. Resources availability
3. Decision environment.

**Pillars - what they are and what they encompass (slide#11-xx)**

The four pillars address (Thomas, 2006):

1. Policies, Standards and Strategy

2. Information Quality

3. Privacy, Compliance and Security

4. Architecture and Integration

The four pillars offer a framework helping companies to develop Data Governance programs to a) achieve clarity

b) ensure value from the efforts

c) create a clear mission

d) maintain scope and focus

e) establish accountabilities

f) define measurable successes.

Each pillar is designed to operate in a cohesive manner where all effort ties back to the universal goals setup to achieve by a company.

One or more pillars should be the focus when developing a Data Governance program. The focus of each pillar will differ depending on the type of rules and issues to be addressed, the emphasis given to certain data-related decision and actions, and level of involvement.

**The value of information/data assets**

Laney (2015) puts 4 key challenges that exist for organizations to effectively measure the value of their information assets.

1. Executives reveal that their organization mismanages its information assets, particularly when compared to how traditional assets are managed leading to missed business opportunities and undue expense.

2. Currently, the accounting profession does not recognize information as a balance sheet asset which means few organizations have a true sense of the value information generates.

3. Research has shown that information-centric companies tend to outperform their peers and that the financial markets tend to favor.

4. Most IT leaders have difficulty demonstrating the economic benefit of key information management initiatives. Such areas include MDM, information governance, information architecture, or business intelligence or analytics.

**Slide #11**

Data as a ubiquitous corporate asset

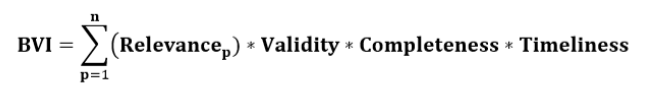
Drive organizational performance

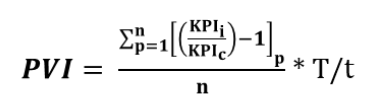
Decision making

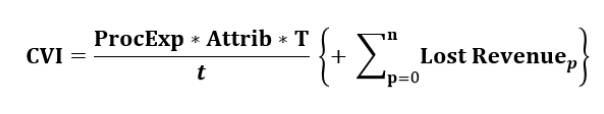
Bottom line profits

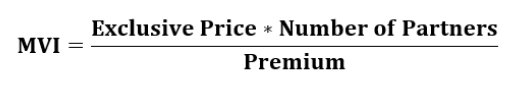
Competitive advantage

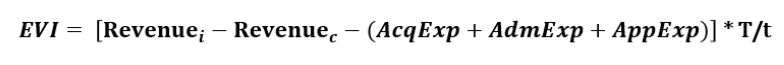










Lecture#2

**When did data protections start to appear in the US**

As early as 1973, government agencies were concerned about the harmful consequences of automated systems that collect and store data. In that year, the Department of Health, Education and Welfare published a report entitled Computers and Rights of Citizens (Cobb, 2016).

**Federal Consumer protection law**

The *Federal Trade Commission (FTC) Act* is the primary statute of the Commission. Under this Act, as amended, the Commission is empowered, among other things, to

1. Prevent unfair methods of competition and unfair or deceptive acts or practices in or affecting commerce
2. Seek monetary redress and other relief for conduct injurious to consumers
3. Prescribe rules defining with specificity acts or practices that are unfair or deceptive, and establishing requirements designed to prevent such acts or practices
4. Gather and compile information and conduct investigations relating to the organization, business, practices, and management of entities engaged in commerce
5. Make reports and legislative recommendations to Congress and the public.

A number of other statutes listed here are enforced under the FTC Act.

<https://www.ftc.gov/enforcement/statutes/federal-trade-commission-act>

This Act's purpose is to safeguard the personal information of consumers. It also prohibits unfair or deceptive acts when collecting the data. The following categories of data are protected: financial and health data, data about children, social security numbers and precise geographic location information. A company that fails to protect consumer data, changes its privacy policies without adequate notice, or fails to comply with a published privacy policy can face prosecution under this Act.

**The Fourth Amendment**

The Fourth Amendment originally enforced the notion that “each man’s home is his castle”, secure from unreasonable searches and seizures of property by the government. It protects against arbitrary arrests, and is the basis of the law regarding search warrants, stop-and-frisk, safety inspections, wiretaps, and other forms of surveillance, as well as being central to many other criminal law topics and to privacy law.

Amendment IV

The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

Lecture#3

**Standards for Security Categorization of Federal Information**

Title III of the E-Government Act, entitled the *Federal Information Security Management Act of 2002 (FISMA)*, tasked *National Institute of Standards and Technology (NIST)* with responsibilities for standards and guidelines, including the development of:

• Standards to be used by all federal agencies to categorize all information and information systems collected or maintained by or on behalf of each agency based on the objectives of providing appropriate levels of information security according to a range of risk levels;

• Guidelines recommending the types of information and information systems to be included in each category; and

• Minimum information security requirements (i.e., management, operational, and technical controls), for information and information systems in each such category. (NIST, 2004)

|  |  |  |  |
| --- | --- | --- | --- |
| **Security Objective** | **LOW** | **MODERATE** | **HIGH** |
| ***Confidentiality***  Preserving authorized restrictions on information **access** and **disclosure**, including means for protecting personal privacy and proprietary information.  [44 U.S.C., SEC. 3542] | The unauthorized disclosure of information could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals. | The unauthorized disclosure of information could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals. | The unauthorized disclosure of information could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals. |
| ***Integrity***  Guarding against improper  information **modification**  or **destruction**, and includes ensuring information **non-repudiation** and **authenticity**.  [44 U.S.C., SEC. 3542] | The unauthorized modification or destruction of information could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals. | The unauthorized modification or destruction of information could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals. | The unauthorized modification or destruction of information could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals. |
| ***Availability***  Ensuring timely and **reliable access** to and **use** of information.  [44 U.S.C., SEC. 3542] | The disruption of access to or use of information or an information system could be expected to have a **limited** adverse effect on organizational operations, organizational assets, or individuals. | The disruption of access to or use of information or an information system could be expected to have a **serious** adverse effect on organizational operations, organizational assets, or individuals. | The disruption of access to or use of information or an information system could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, or individuals. |

**PII vs PHI (slide#67-68)**

PHI is an acronym of Protected Health Information, while PII is an acronym of Personally Identifiable Information. ... Personally identifiable information (PII) or individually identifiable health information (IIHI) is any health information that allows the patient to be identified.

HIPAA uses the term Protected Health Information (PHI) to refer to protected data, but the concept is very similar to the term Personally Identifiable Information (PII), which is used in other compliance regimes. ... Certain information like full name, date of birth, address and biometric data are always considered PII.

**Controlled Technical Information ??**

Cross-platform policy capabilities will be key to meeting enterprises’ need for securing sensitive big

data, as will the following technical tactics:

* Fine-grained user access control. Using roles and user-specific or general attributes, such as

location, is a must for any organization that wants to deploy data security best practices such as

* limiting access to enterprise data based on a least-privilege model or on a need-to-know basis.
* Support for row, column, and cell-level data access policies. The ability to apply policies down to the
* most granular element of data provides organizations with the greatest degree of control and flexibility
* over who has access to which specific data.
* Granular and dynamic data masking. Early adopters of Hadoop have gone as far as duplicating
* datasets to enable masking capabilities. Advanced security approaches for big data environments
* should deliver masking policies that support the same level of granularity as user and data access
* control policies without the need for data duplication.
* Ability to use data without revealing it. This capability enables enterprises to use sensitive data in
* the query logic of their applications without actually revealing the data itself to end users.
* Data access auditing. The ability to log and track which user has access to which data will simplify
* the process of auditing for various compliance requirements.
* Vendor neutrality. A data access solution should perform consistently across different Hadoop
* vendors (Cloudera, Hortonworks, MapR) and SQL data stores with support for cloud deployments.

<https://drive.google.com/file/d/1zG_CZEErIwb2VUfPHOz1WC0ZZJ_Wgpo9/view?pli=1>

Lecture#4

**Data Mining (slide 98-101)**

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.

Nielsen (2015), in the article “Who Owns Big Data?” brings up an interesting dilemma. The CEO of Google in 2010 realized that by analyzing all the queries of its users, it can determine a collective intent, and could use that to predict the stock market. Allegedly, it did not continue down that path, over concerns of legality. It did proceed to create a database of Flu trends, which tracks influenza across the country.

Facebook has created a program that predicts which Facebook users will end up in relationships.

Other researches have used Twitter to predict box office revenues.

All of these are available because a small team of programmers have access to an avenue of big data that is not available to the general public. Should there be a big data database open to the public? Nielsen opines that there should. The government has begun this process through its website data.gov. However, this is limited to government data. Aggregating data from multiple sources would be the most beneficial and spur innovation for those without access to large company databases.

**Electronic Evidence (related to email)**

**The Electronic Communications Privacy Act (18 U.S.C. §2510) (ECPA)**

Emails on the server (after 180 days) and personal computers (requires warrant; regardless how old)

**Data Map (Data Lineage)**

In any situation involving litigation or security, an organization needs to be able to locate and document the existence of electronic data. Data maps are essential to this task. This involves:

Documenting the databases  
Understanding what they contain  
Knowing the department that controls the data within them  
Contact information of the person within that department that has actual control of the data  
The person who can authorize the release of the data

Sample of a Data Map

(Mathews, 2016, p. 169)

Lecture#5

**Moral Agency (slide 141-142)**

Moral agency is an individual's ability to make moral judgments based on some notion of right and wrong and to be held accountable for these actions. A moral agent is "a being who is capable of acting with reference to right and wrong."

A moral agent is a person who has the ability to discern right from wrong and to be held accountable for his or her own actions. Moral agents have a moral responsibility not to cause unjustified harm. Traditionally, moral agency is assigned only to those who can be held responsible for their actions.

**Managing Access to data (slide#137)**

As we have learned in the previous weeks, classifying data is critical in a data management plan. Once the data is classified, it must be secured. Here is where risk management techniques come into play. The higher the risk if the data is compromised, the higher the level of security. The important element here are the controls surrounding who can access what data. As discussed in Matthews (2016), “Electronically stored information: The complete guide to management, understanding, acquisition, storage, search, and retrieval,” Chapter 6, there are several ways to control access: authentication, encryption, identify management, logical compartmentalization and physical separation, and access control.

Authentication Methods

Authentication methods are the **most common** form of controlling access. This is either something known, like a username and password (something you have), like a smart card; or biometric (something you are).

We are seeing more **two-factor authentication** these days, where two forms of authentication are required.

Encryption

Encryption may be **the most secure method** of access control; however, it is also the most prone to **user inconvenience** and inefficiency. A simple form of encryption is password protection. A more complex form involves lengthy encryption keys which require storage and potential digital logistical nightmares.

Identity Management

Making sure the person on the other end of the request for access, is indeed the correct person is a key step in any access control process.

Having **a single sign-on** that allows a user access to all authorized data is a much more efficient and appreciated approach, from the user's perspective. **User convenience** is usually at odds with security. While single sign-ons are definitely convenient, they provide a single point of resistance for someone trying to access the system. This makes it **inappropriate for highly sensitive data**.

Compartmentalization

Disk partitions are another method of controlling access. Rights and privileges are assigned depending on who has access to what partition.

e.g. Protected shared files with access control as per role.

Physical separation

Using a physical barrier, such as a locked room, is yet another method of controlling access. Users would need to gain access to a secured room in order to gain access to the servers and data that were located there.

**Factors of Authentication**

Look authentication method above

Lecture#6

**Authoritative data (slide#172)**

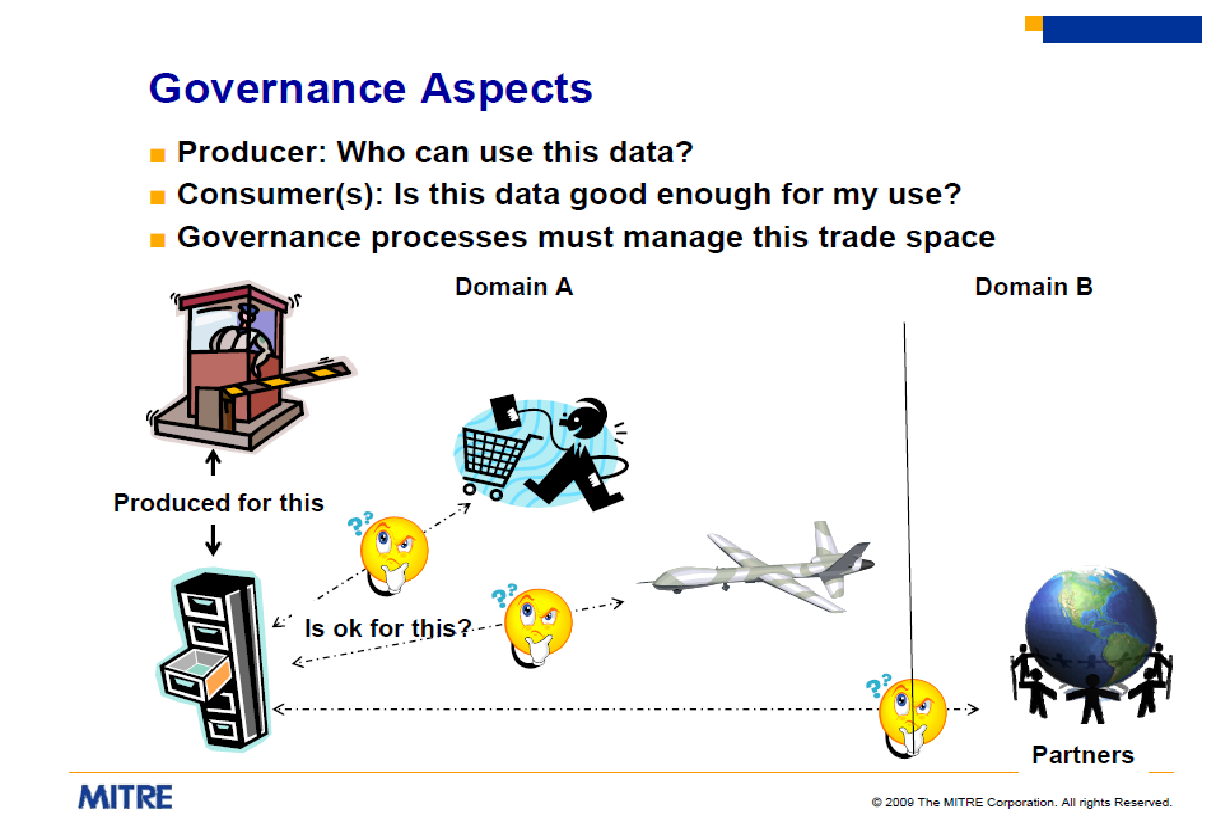
**Authoritative data governance**

Data Governance and Stewardship to drive authoritative data (Westman, 2009)

Authoritative data sources are recognized or official data with a designated mission statement to publish reliable and accurate data for subsequent use by customers. An authoritative data source may be the functional combination of multiple, separate data sources. This type data sources comprise a repository of valid or trusted data “that is recognized by an appropriate 4

Data Governance, Laws & Ethics Week 6 set of governance entities and supports the governance entity’s business environment.” Each governance entity establishes its criteria in the following areas, which may vary per business environment, subset of operations within the business environment, and by Authoritative Source.

* Data that needs to be collected
* Data that is collected
* Data quality (accuracy, reliability, freshness, etc.)
* Data usage (aka what data can be used for)
* Assurance requirements
* Compliance requirements



**AIMQ**

Techniques for assessing information quality

AIMQ (Lee et al. 2002)

Lee et al. (2002) developed the AIM quality (AIMQ) methodology for assessing and benchmarking Information Quality (IQ) in organizations. This methodology encompasses three major components: the PSP/IQ model, the IQA instrument, and the Gap Analysis techniques. Each component of the AIMQ has merit in itself and, therefore, makes a contribution on its own.

The PSP/IQ model assesses IQ in terms of conformance to specifications and as exceeding consumer expectations on the one axis and IQ as a product and as a service on the other. It is a management tool for conceptualizing and assessing IQ in business terms. Also, the model organizes the key IQ dimensions so that meaningful decisions can be made about improving IQ. Furthermore, it serves as a theoretical foundation for performing gap analyses.

The IQA instrument provides the measurements underpinning the PSP/IQ model and the gap analyses. It collects data to assess IQ status along the key IQ dimensions. The data collected from the IQA are the prerequisite for PSP/IQ modeling and gap analyses. The gap analysis techniques provide the tools by which organizations can understand their IQ deficiencies as compared to other organizations and to different stakeholders within one organization. Using these analysis techniques, organizations can benchmark their IQ and determine appropriate areas to focus improvement efforts.

The AIMQ methodology as a whole provides a practical IQ tool to organizations. It has been applied in various organizational settings, such as the financial, healthcare, and manufacturing industries. The methodology is also useful in identifying IQ problems, prioritizing areas for IQ improvement, and monitoring IQ improvements over time.

**Data Quality/Information Quality participants (Slide#175)**

Data quality is the ability of a given data set to serve an intended purpose. Maintaining data quality is a difficult but necessary task. In order to achieve consistent and reliable customer data, businesses must constantly manage data quality so that they can trust their data and enable quicker and more knowledgeable decisions.

Key roles in organizations to support stewardship include:

**Data Architects** - The Data Architect is responsible for the data model and making it business relevant as well as establishing and maintaining naming standards. The Data Architect is responsible for managing and disseminating metadata to facilitate the common understanding of data and encouraging its reuse.

**Data Suppliers** - The Data Supplier creates data and maintains its consistency and correctness.

Lecture#7

**Cybersecurity legislation (slide#234-235)**

Cybersecurity Information Sharing Act of 2015 (US Congress, 2015)

U.S. Congress passed this bill to improve cybersecurity in the United States through enhanced sharing of information about cybersecurity threats, and for other purposes. This bill requires the Director of National Intelligence and the Departments of Homeland Security (DHS), Defense, and Justice to develop procedures to share cybersecurity threat information with private entities, nonfederal government agencies, state, tribal, and local governments, the public, and entities under threats.

**Information governance tools and vendors**

Third-party vendors (De Simoni and Edjlali, 2016)

“The growing need for organizations to treat information asset as an asset is making metadata management strategic, driving significant growth. By 2020, 50% of information governance initiatives will be enacted with policies based on metadata alone. Vendors in this market should demonstrate the ability to identify, document, and maintain relationships among ingested and translated metadata. Niche players are Data Advantage Group, SAP, Cambridge Semantics. Visionaries vendors are Informatica, Collibra, IBM, Adaptive, Global IDs, an Oracle.”

Key functionality sought out in products are:

* Metadata repositories
* Business glossary
* Data linage
* Impact analysis
* Rules management
* Semantic frameworks
* Metadata ingestion and translation
* Database management system catalogs
* ERP and other applications
* XML formats
* Hardware and network log files
* Spreadsheets and Word document
* PDF documents
* Business meta
* Custom Metadata

**Semantic Technology (and ontology) – (Slide#222 and 227) – Slide#210**

Organizations at the forefront of their respective industries have started to extend their enterprise architectures with semantic technologies.

The internet services as a model for Semantic technologies as the methods and tools originate from World Wide Web Consortium (W3C) and have been adopted for enterprise needs (Sabou, Ekaputra, and Biffl, 2017). Semantic Web - W3C

Leading companies from sectors including healthcare and pharmaceutical, finance, media and publishing, and ecommerce, are establishing knowledge modeling and graph-based approaches at the core of effective data management. OWL - Semantic Web Standards

Semantic software solutions give data scientists and subject-matter experts broad leeway for structured and unstructured data operations across platforms.

Together with enterprise architects, these specialists are shaping highly customized digital environments for their organizations.

<https://www.marklogic.com/blog/eliminate-data-silos-multi-model-database/>

<https://drive.google.com/file/d/1biMIs2uNJNAlxWY-SDH2Uc4Cen73cZZe/view?pli=1>