Lecture Notes
Week 08

INFS3200 Advanced Database Systems
Semester 1, 2021

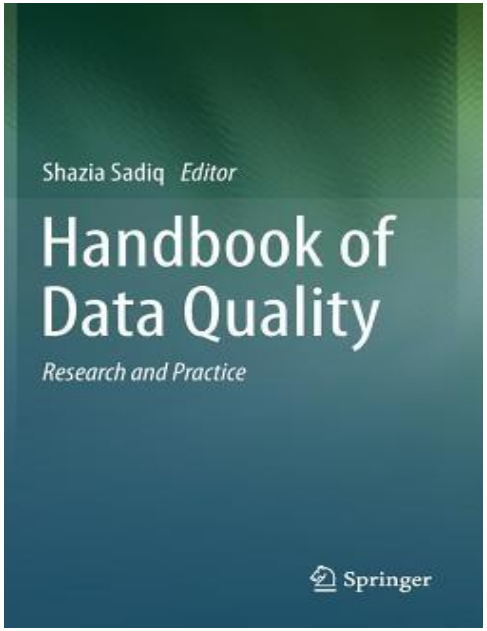
Data Quality Management

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Professor Xue Li

+ Outline

- Data Quality Dimensions
  - Concept of data quality
  - Data life cycle
- Management of Data quality
  - Measurement of data quality costs and improvements
  - Strategies for data quality improvement
  - Maturity models and data governance
  - Computational approaches
- Four basic Steps of Data Governance



What is data?

Happy !

Yahoo !

18

Data is an accurate description of facts

What is Data?

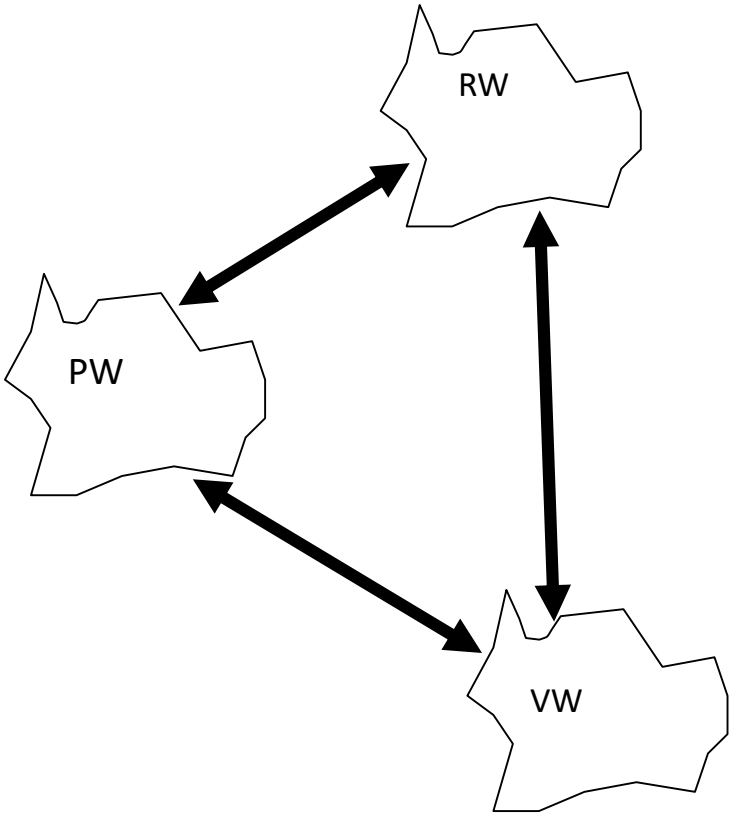
Four properties of data:

1. **Meta Data & Data: Two levels of meaning -** it's the *description* and *measurement* about objects (such as patient records, observed temperature values, financial records, etc.
2. **Constraints: Associated context -** the time it is collected, the constraints imposed on it (data domain: ranges, conditions, cardinality, precision, etc), the format (e.g., inch vs, cm, ponds vs Kg, ets), etc.
3. **Data Structures: Associated Relationship Structures -** dependency, causality, co-occurrence, covariance, contra-variance, association, correlation, **dimensionality**, ratio (1:m, m:n, 1:1), etc.
4. **Dynamics: Associated Changes -** monotonous changes, patterned changes, etc.

In a view of computer, data is an organized storage of binary numbers.

The Three Worlds Vision

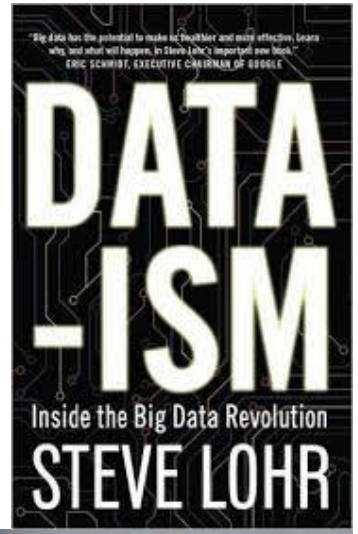
- **Real World** (RW)
  - Physical (natural world)
  - Objective
- **Virtual World** (VW)
  - Digital (Internet)
  - Reflective
- **Perceptual World** (PW)
  - Conceptual (human brain)
  - subjective



Data-ism: What connections are there in between? How do we map them?

Data-ism

- **High Dimensionality：** Every piece of data is **a point** in a high-dimensional data space.
- **Sparsity：** In a Problem Space, the Data Space is always **sparse**.
- **Uniqueness：** In a high-dimensional space, every data point is an **outlier**.
- **Relevant Connectivity:** Data is about relationships, data is understood as a graph (network), data is expressed as matrixes.

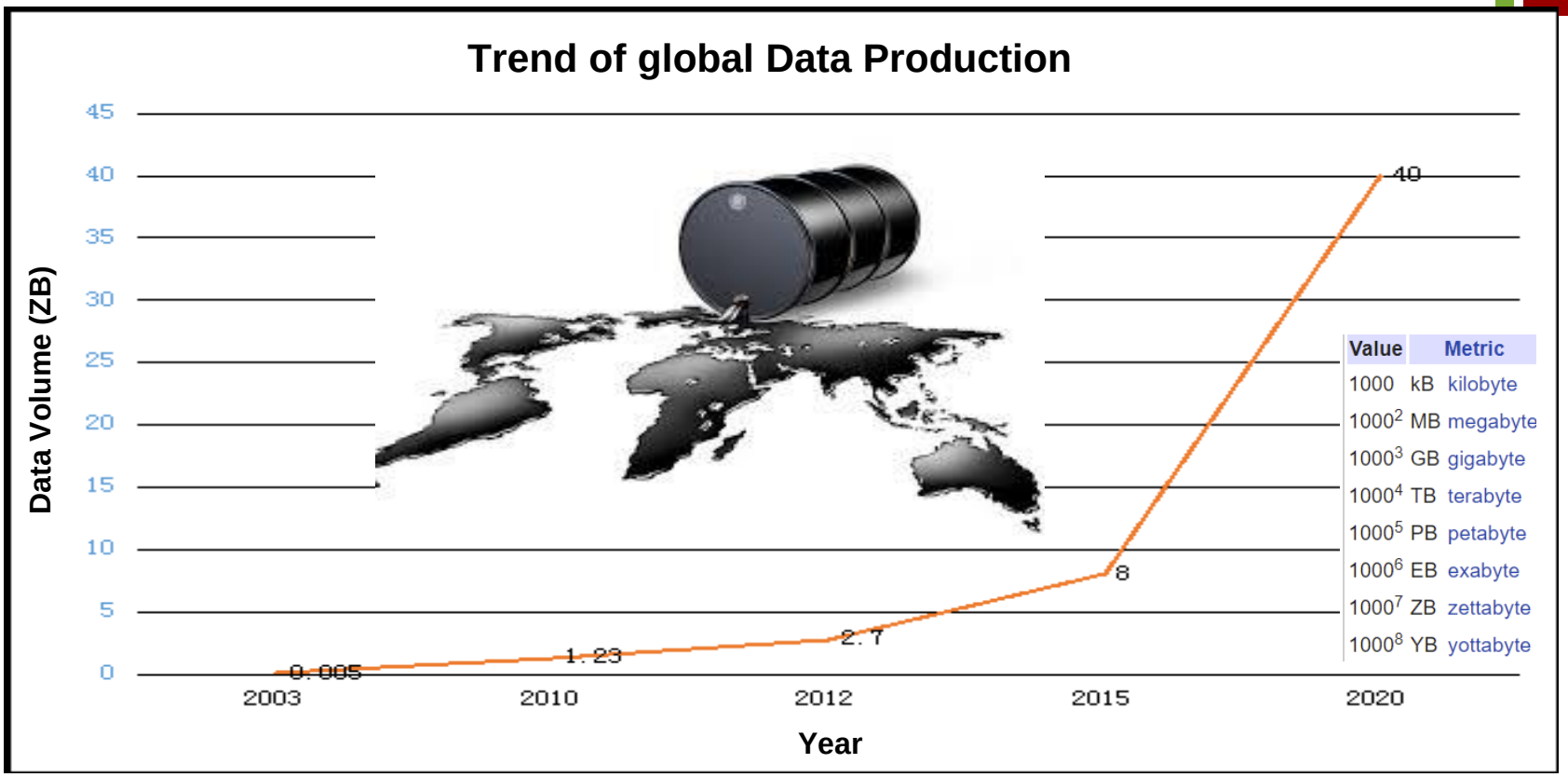


An era of data is coming!

Challenges:
Development of Hardware, Data, & Knowledge

+ “Data is the new oil of industry.”

- Dr Leif Hanlen, Technology Director, DATA61.



http://www.09dsj.com/index.php/archives/6515

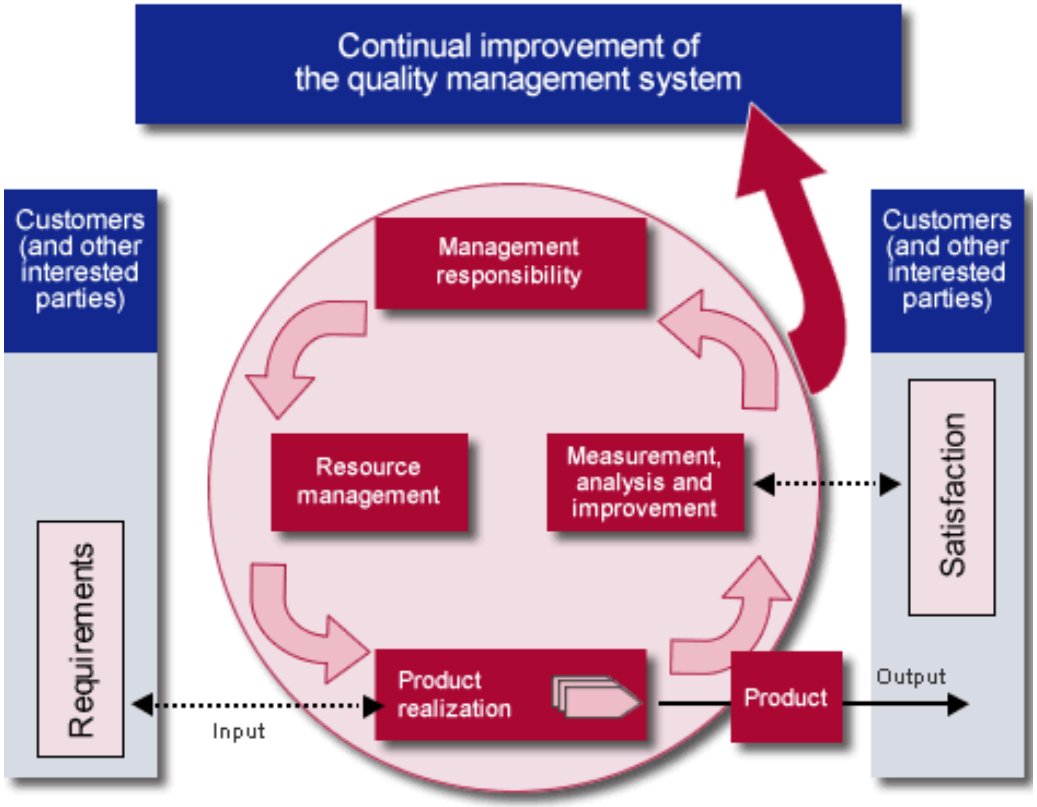
“Business is now all about Data”

“Agriculture is about data, health is about data, finance is about data, and sport is about data.”

- Prof Hugh Durrant-Whyte, Sydney University

What is Quality?

- ISO 9000 is a family of standards for quality management systems.



ISO9000 Process Approach

- Although the standards originated in manufacturing, they are now used in organizations that produce various types of products which include a **physical object,** or **services,** or **software**.
- It is the world's most established quality framework, currently being used by over millions of organizations in 161 countries.

www.iso.org

+ Governance vs. Management

- "**Governance**" is the strategic task of setting the organisation's goals, direction, limitations and accountability frameworks.
- **"Management"** is the allocation of resources and overseeing the day-to-day operations of the organisation.
- One way to think about this is that
  - **Governance** determines the "What?" - what the organisation does and what it should become in the future.
  - **Management** determines the "How?" - how the organisation will reach those goals and aspirations.

https://www.researchgate.net/post/What_is_the_difference_between_governance_and_management#:~:text=%22Governance%22%20is%20the%20strategic%20task,day%20operations%20of%20the%20organisation.

Francesco Appio
SKEMA Business School

Information System Quality

- (Computer) System Quality
  - Throughput, reliability, availability, response time, resilience
- Method Quality
  - Software Quality
    - Coupling, cohesion, complexity, modularity, size…
  - Model Quality
    - Correctness, completeness, understandability, compactness, precision (no ambiguity)…
- Data Quality

DeLone, W.H., and McLean, E.R. Information systems success: The quest for the dependent variable. *Information Systems Research, 3*, 1 (1992), 60-95.

+ Data Quality Issues: Example

misleading

inaccurate

inappropriate key?

duplicate

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success	Information Systems Research	3	1	1996	Pp 60-95
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

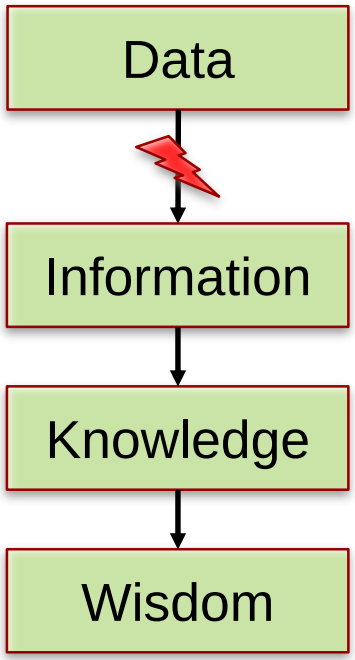
Authors? incomplete

incomplete

invalid

+ Impact of Data Quality

- Customer Dissatisfaction
- High Costs
- Low Job Satisfaction
- Organizational Mistrust
- Poor Decision Making
- Impedes Re-engineering
- Hinders Long-term Business Strategy



→ Breaking the Information Chain

D. Tapscott (1996) The Digital Economy: Promise and Peril in the Age of Networked Intelligence, New York: McGraw-Hill, 1996.

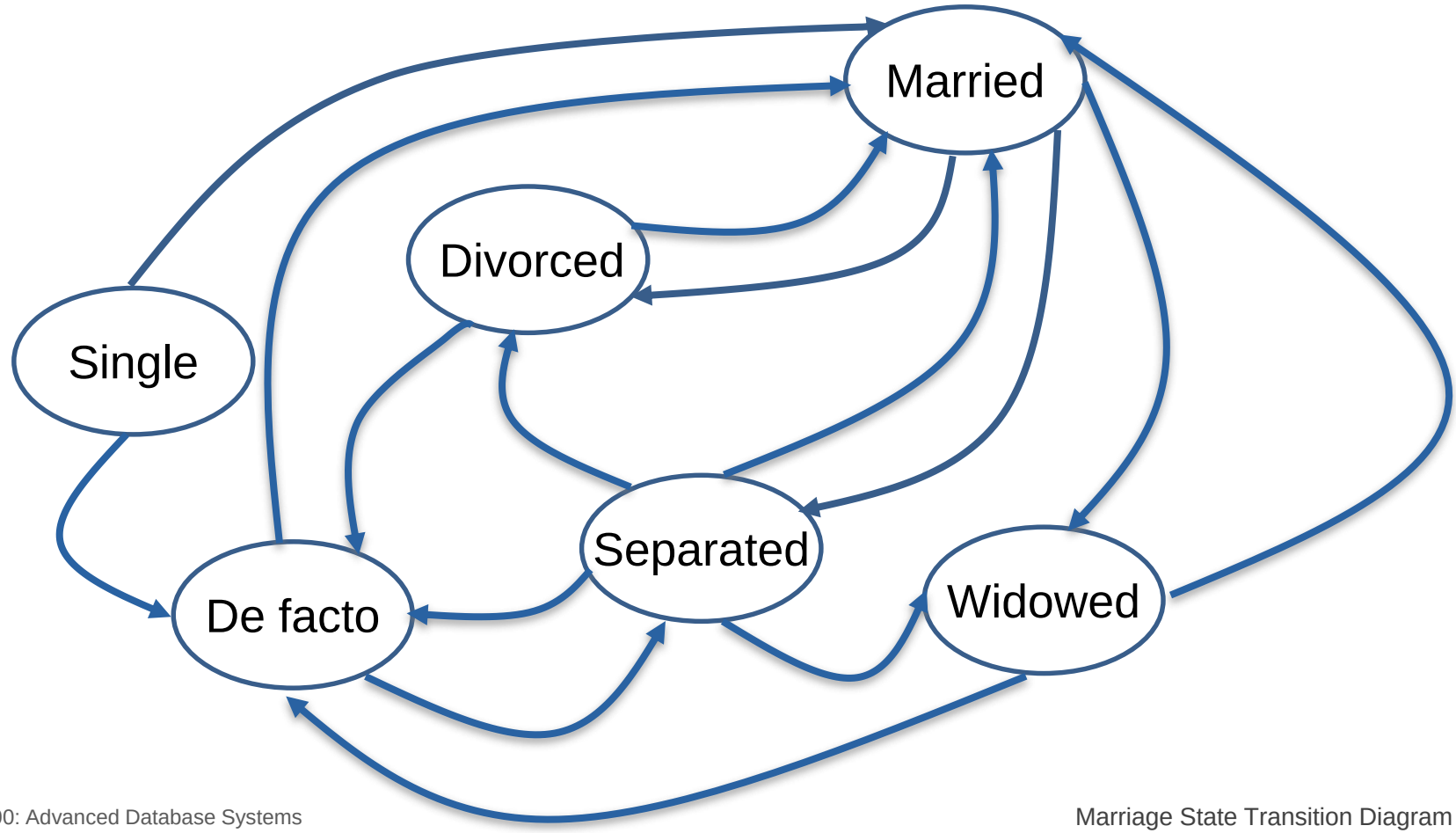
Redman (1996) Data Quality for the Information age. Artech House, Norwood, MA., USA. 1996.

Data Quality Dimensions

- Integrity (Meaningless)
- Accuracy (Erroneous)
  - Postcode “4109” is typed “4019”
- Completeness (Missing)
  - Students don’t have to declare a major till graduation, so major is missing in most enrolments
- Currency (Obsolete)
  - Old phone numbers
- Representational Consistency (Inconsistent)
  - ITEE Vs. Information Technology and Electrical Engineering
- Accessibility (Unavailable)
  - Server down, privacy concerns
- Reliability & Trust (Uncertainty)

+ Integrity

Question: Can we modify one's marriage status in a database without considering this pattern?



Marriage State Transition Diagram

+ Integrity Constraints

- Good DB design allows RDBMS to enforce integrity constraints ... but cannot be taken for granted in real life
  - Key dependency (uniqueness constraint)
  - Inclusion dependency (referential integrity)
  - Functional dependency
  - Domain constraints
  - Dynamic vs Static constraints
- Database Constraints vs. Business Application Constraints
  - a person's *Salary* cannot be reduced
  - *Weekly_Income* must be updated together with the *Repay_Amount*
  - *Expire_Date > Issue_Date*
  - Deposit of the Money can only be performed between the time 12-3 pm, Monday to Friday
  - *SUM(DEPT.Salary) < BUDGET.Salary*

+ Accuracy

- "Accuracy is defined as the closeness between a value *v* and a value *v'* considered as the correct representation of the phenomenon that *v* aims to represent"

Inaccurate (should be 1996)

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

invalid

Carlo Batini, Monica Scannapieco (2006) Data Quality Concepts, Methods and Techniques, Springer 2006.

+ Completeness

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

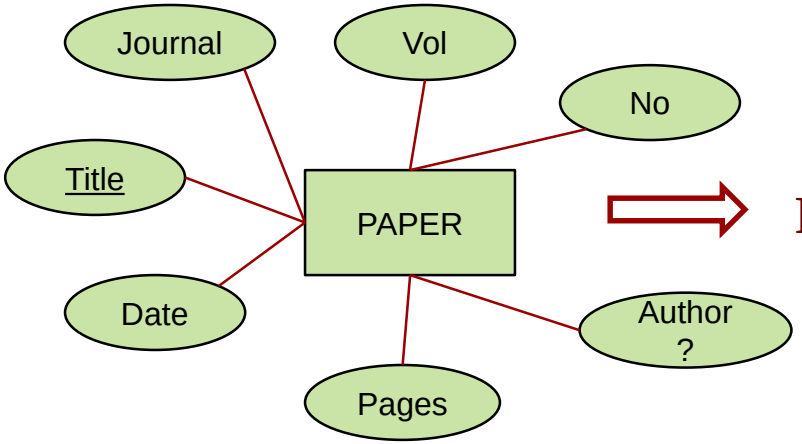
Authors? incomplete

incomplete

- Completeness is defined as the sufficiency of data for the task at hand

+ Schema Completeness

- The degree to which concepts are not missing from the schema



PAPER(Title, Journal, Vol, No, Date, Pages)

+ Column Completeness

- Measure of missing values in a column (tuple)
  - Does not exist (Driving licence no. for a child)
  - Exists but missing (height of a person)
  - Unknown if exists (email of a person)

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

+ Population Completeness

- Missing values with respect to a reference
- The number of registered residents in the city council of Springfield are 6,000,000
- Company “Annoying TeleMarketing” keeps a record of residents and has a current population (r) of 5,400,000
- Population Completeness C(r) = |r| / |ref(r)| = 0.9

+ Completeness and the World

	Closed World Assumption	Open World Assumption
NULLS present	Data is known to be incomplete and incompleteness can be measured	Data is known to be incomplete and incompleteness cannot be measured accurately
NULLS absent	Data is complete	Data can be incomplete but it is unknown what is missing ... need a reference table

+ Currency

- Currency
  - How promptly data is updated
- Volatility
  - Frequency with which the data vary in time
- Timeliness
  - How current the data are for the task at hand

+ Consistency

- Lack of consistency is the result of violation of integrity constraints

inappropriate key?

duplicate

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success	Information Systems Research	3	1	1996	Pp 60-95
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

+ Presentation Issues

- Interface design
- Data entry validation
- Training of operators

Misleading YEAR?

Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

+ Impact of Presentation

- Date of Payment
  - What does this mean? when payment was authorized, when payment request was submitted, when the cheque was made, when cheque was posted, when payment was received, when payment date was entered in the system?
- “Dear Deceased”
  - A home insurance clerk put “Deceased” in the name of a customer who was co-owner of the insurance policy. When the next letter was sent it said” Dear Mr. Deceased and Mrs. Smith”

+ Beyond the Content

- Accessibility
  - Timing & Availability (24/7)
  - Server down
- Reliability
  - Trust, Believability, Reputation, ...
- Volume of data
  - Value derived
  - Interpretability
  - Ease of understanding

+ Many More DQ Dimensions

Category	Dimension	Definition: the extent to which
Intrinsic	Believability	Data are accepted or regarded true, real and credible
	Accuracy	Data are correct, reliable and certified free of error
	Objectivity	Data are unbiased and impartial
	Reputation	Data are trusted in terms of source and content
Contextual	Value-added	Data are beneficial and provide advantage for their use
	Relevancy	Data are applicable and useful for the task at hand
	Timeliness	The age of the data is appropriate for the task at hand
	Completeness	Data are sufficient for the task at hand
	Appropriate amount of data	The quantity or volume of available data is appropriate
Representational	Interpretability	Data are in appropriate language and unit
	Ease of understanding	Data are clear without ambiguity and easily comprehended
	Representational consistency	Data are always presented in same format
	Concise representation	Data are compactly represented
Accessibility	Accessibility	Data are available or quickly retrieved
	Access security	Access to data can be restricted and hence kept secure

Wang & Strong (1996) Beyond Accuracy: what data Quality means to Data Consumers. Journal of MIS 12/4. 1996

+ Data Quality: Summary

- Quality control is a profession which spans products and manufacturing, management practices, services, software, and all kinds of systems
  - Data Quality is an aspect of information system quality
  - Data Quality is not only about (raw) data
  - Data Quality is a multi-dimensional problem
  - Data Quality is sometimes subjective

+ Data Quality Management

- Different perspectives
- Problems and solutions
- Data quality governance

+ Historical Perspective

- "Data as a product"
- Assessment of data quality through an analogy with product quality standards (ISO9000)

	Product Manufacturing	Data Manufacturing
Input	Raw Material	Raw Data
Process	Materials Processing	Data Processing
Output	Physical Products	Data Products

Data ***is*** the product ... Intention and usage of data is no longer aligned in current large scale information systems.

Wang, Storey & Firth (1995) A Framework for Data Quality research. IEEE TKDE Vol.7/4, Aug, 1995.

+ Application Perspective

- Data Quality Rules based on "Use it or lose it"
  - DQ1: Unused data cannot remain correct for very long
  - DQ2: Data quality is a function of its use, not its collection
  - DQ3: Data quality will ultimately be no better than its most stringent use
  - DQ4: The less likely some data attribute is to change, the more traumatic it will be when it finally does change
  - DQ5: Data Quality problems tend to become worse as the system ages
  - DQ6: Laws of data quality apply equally to data and metadata

No absolute data quality metrics ... Diversity of **data usage** makes quality definitions application/usage specific and feedback driven.

Ken Orr. (1998) Data Quality and Systems. Comm. ACM 41/2. Feb 1998

+ Structural Perspective

- The Relational Model can be an effective means to ensure data integrity

Traditional quality assurance techniques e.g integrity constraints, concurrency control etc. fail (or extremely difficult to enforce) in large scale collaborative information systems.

Unstructured and semi-structured data (XML) as well as multi-media data (Video, Maps, Images etc.) dominant in current information systems as is varied forms of data collection (sensor devices, medical instrumentation, RFID)

Data is of different formats and is **shared** ... A central control of data quality is not possible

+ Data Life Cycle: an Old & Static View

Design requirements and usage benefits may not have explicit connections

Data ***is*** the product

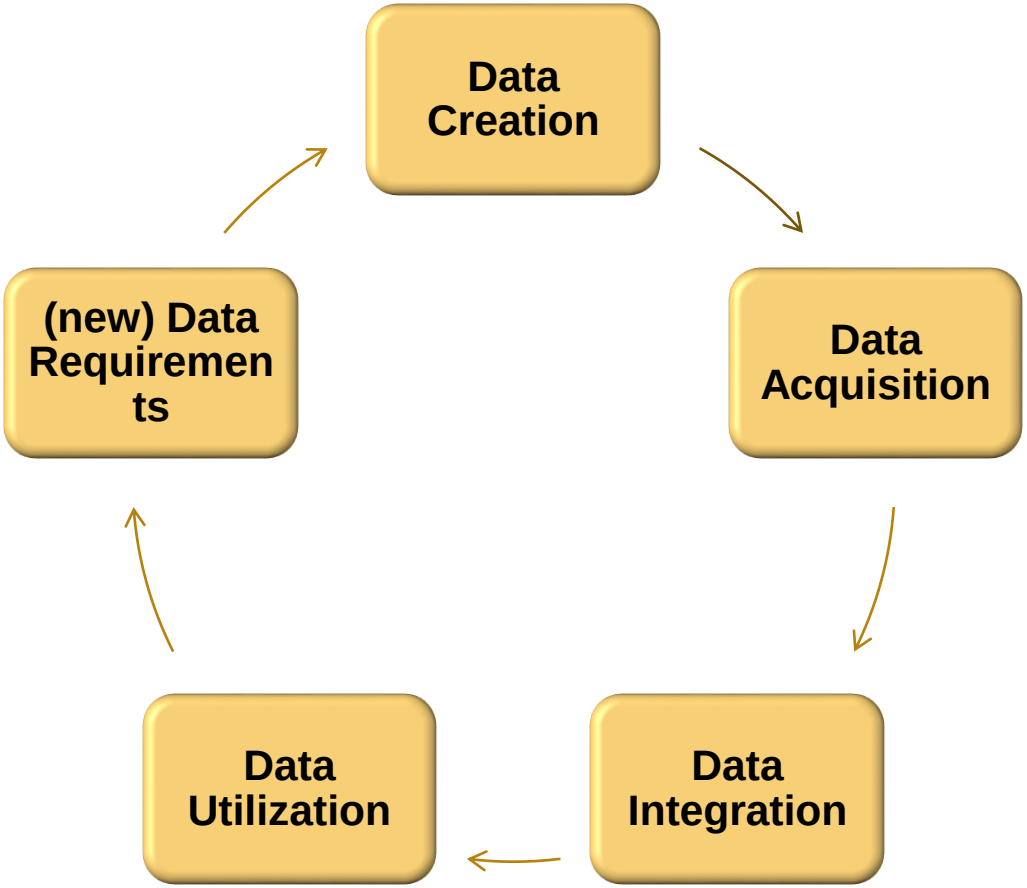
Data is being integrated, more than it is being designed

Data is shared

Diversity in data format, scale and usage unprecedented

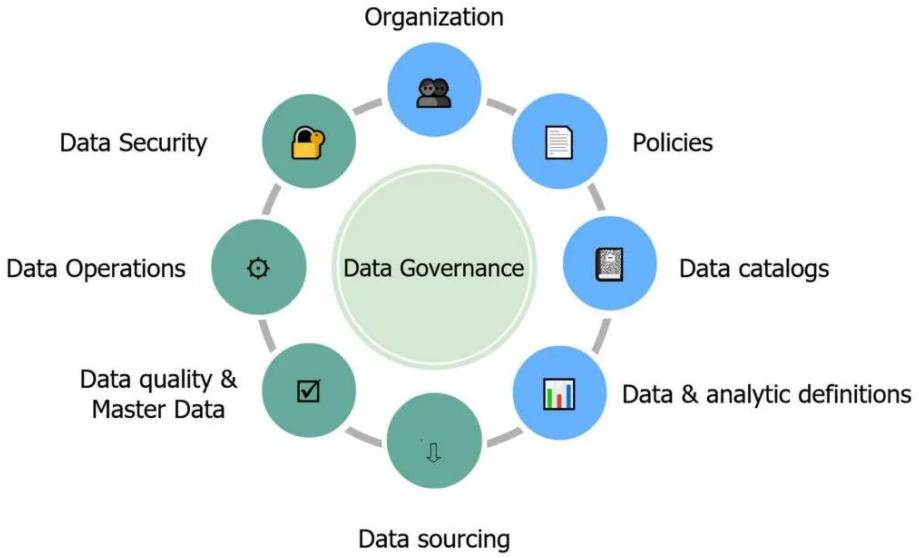
No absolute data quality metrics

+ Data Life Cycle: a New & Dynamic View



+ Four basic Steps of Data Governance (4-1)

- **Recognize the problem**
- Measure its costs
- Devise strategy for improvement
- Aim towards data governance maturity



https://www.imperva.com/learn/data-security/data-governance/

https://www.avenga.com/magazine/data-governance-what-for-why-and-how/

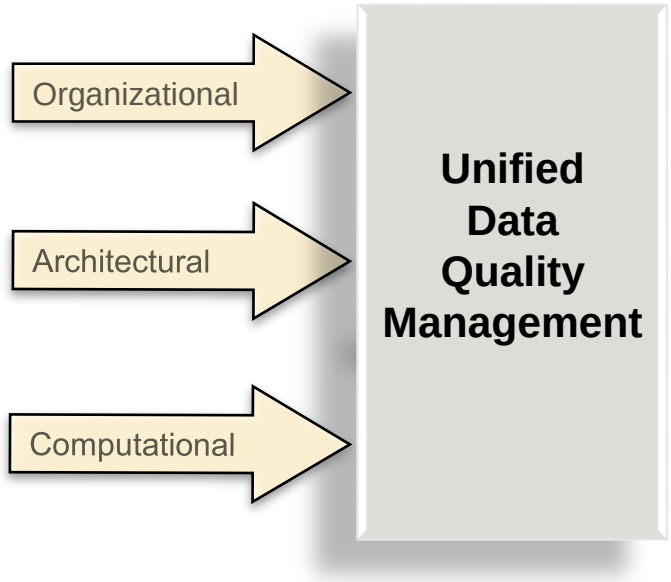
+ Data Quality Problems

Data Quality w.r.t. three aspects:

Data Acquisition	Data Integration	Data Utilization
Data entry errors, poor input validation, lack of user training	Differences in format	Informational overload/ relevance/usability
Character recognition in document management systems	Structural (schema) and semantic (constraints, values) mismatches	Real time analytics, business intelligence
Performance of RFID readers	Privacy preservation, access control	Data mining & statistical (un)truths
Dealing with ambient noise in sensor network communication…	Error propagation	Privacy & trust…
	Attribution, audit, tracking, lineage...	

+ Problems → Solutions

- Who is interested in DQ research & practise?
- Business Analysts
  - Organizational solutions
- Solution Architects
  - Architectural solutions
- Database Experts/ Statisticians
  - Computational solutions

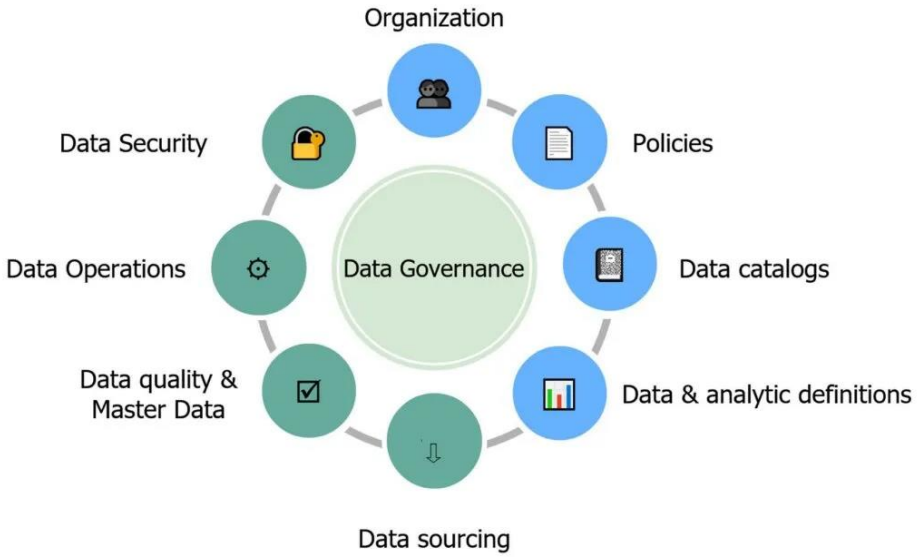


Data Quality Solutions

	Data Acquisition	Data Integration	Data Utilization
Organizational solutions	Development of data quality **objectives** for the organization and **strategies** to establish the people, processes, policies, and standards required to manage and ensure the data quality objectives are met.		
Architectural solutions	The **technology** landscape required to deploy developed data quality management **processes**, **standards** and **policies**.		
Computational solutions	Effective and efficient **tools & techniques** required to meet data quality objectives.		

+ Four basic Steps of Data Governance (4-2)

- Recognize the problem
- **Measure its costs**
- Devise strategy for improvement
- Aim towards data governance maturity

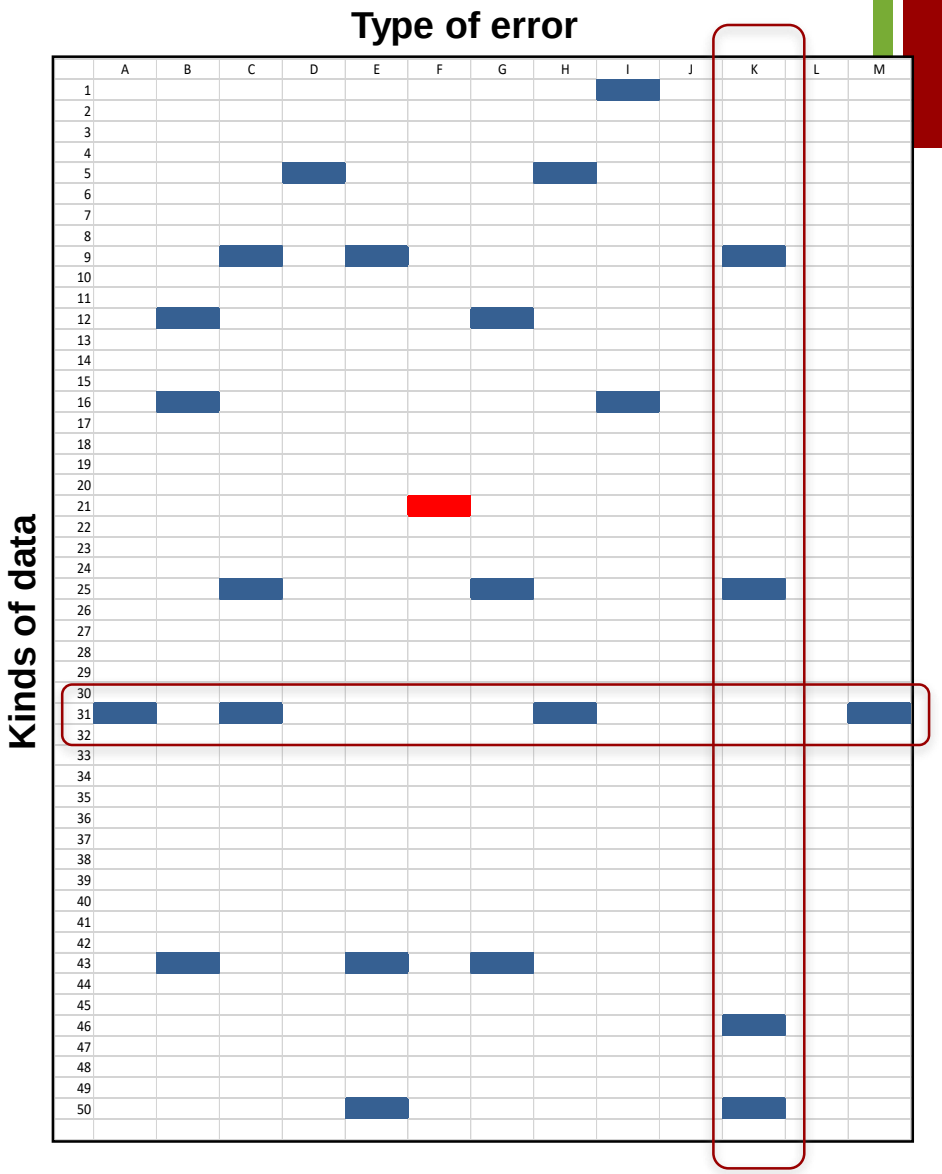


https://www.imperva.com/learn/data-security/data-governance/

https://www.avenga.com/magazine/data-governance-what-for-why-and-how/

Each Error is a Cost

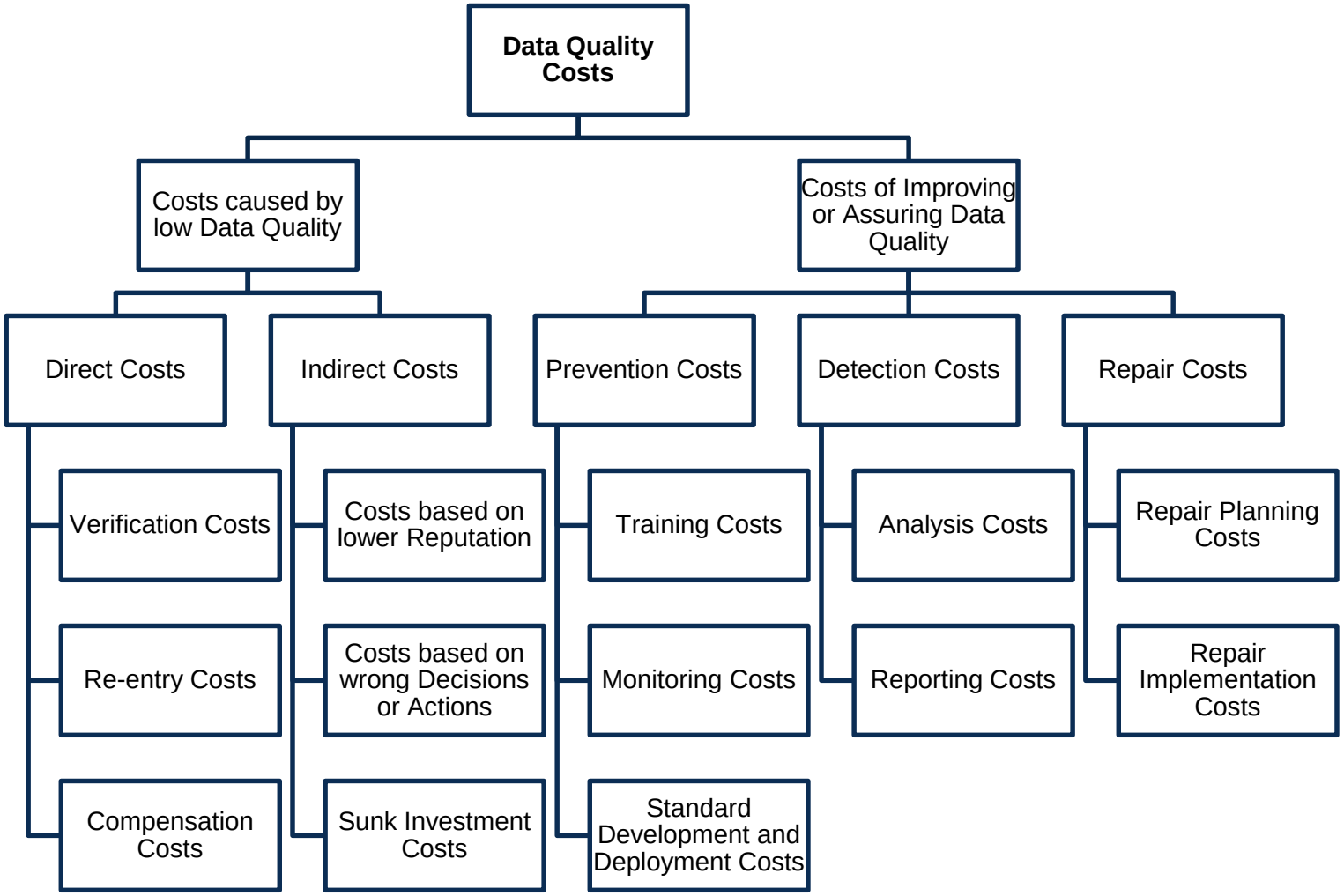
- How much is each error costing the organization?
- How much will the organization save if the errors are eliminated (reduced)



+ Measuring Cost: Example

- Metric: Number of duplicates
- Cost (Benefit): Cost of returned mail e.g. in a marketing campaign
- Success threshold: Reduce number of duplicates by 15%

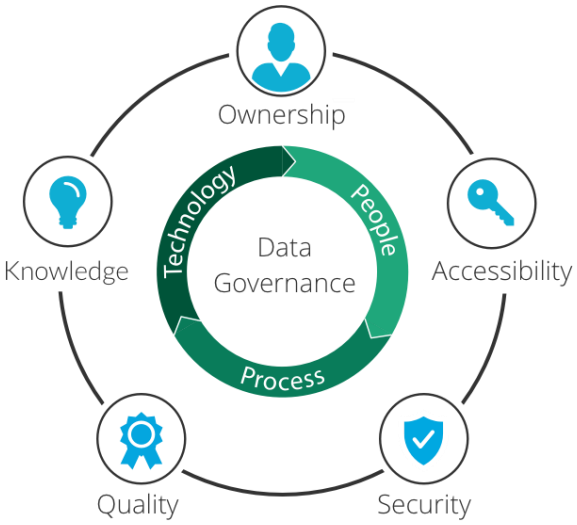
+ Cost/Benefit Analysis



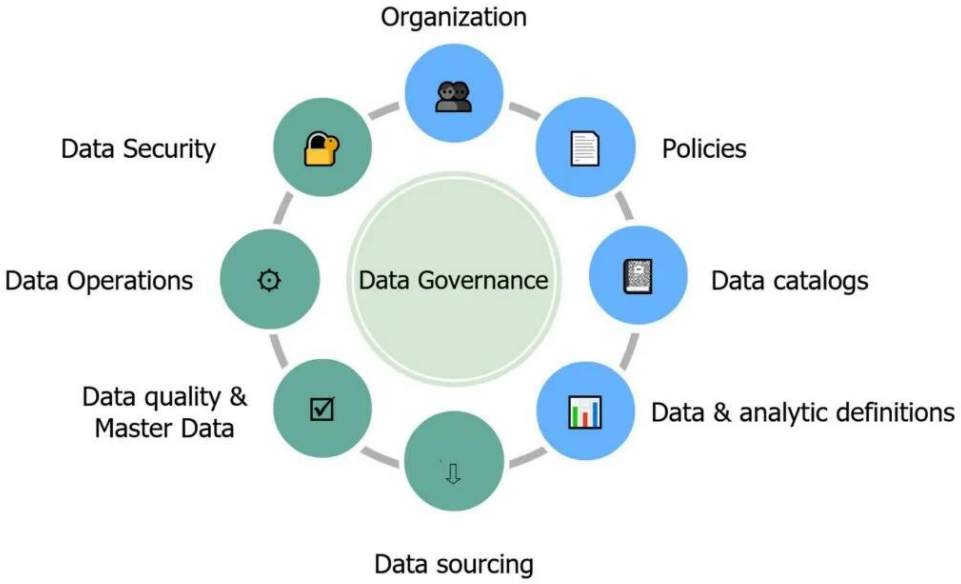
M.J. Eppler, M.A. Helfert (2004) A Classification and Analysis of Data Quality Costs. Proc. 9th International Conference on Information Quality (IQ2004)

+ Four basic Steps of Data Governance (4-3)

- Recognize the problem
- Measure its costs
- **Devise strategy for improvement**
- Aim towards data governance maturity



Data control



https://www.imperva.com/learn/data-security/data-governance/

https://www.avenga.com/magazine/data-governance-what-for-why-and-how/

+ A Sample Strategy

- Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.
- Each of the subsidiaries may have its own partner suppliers along with item catalogs.
- Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.
- However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a **reference (synonym) table** for suppliers
2. **Load supplier da**ta from all subsidiaries into the reference table
3. Use *matching* techniques to **identify potential overlaps**
4. Extract a **master table** for suppliers – represents a single version of truth
5. Retain original representations – represent **multiple versions** of truth
6. Allow access for subsidiaries to reference **master data in all new** (or update) transactions involving supplier data
7. Ensure data managers are **accountable** for continued master data checks
8. Introduce a periodic **monitoring** scheme

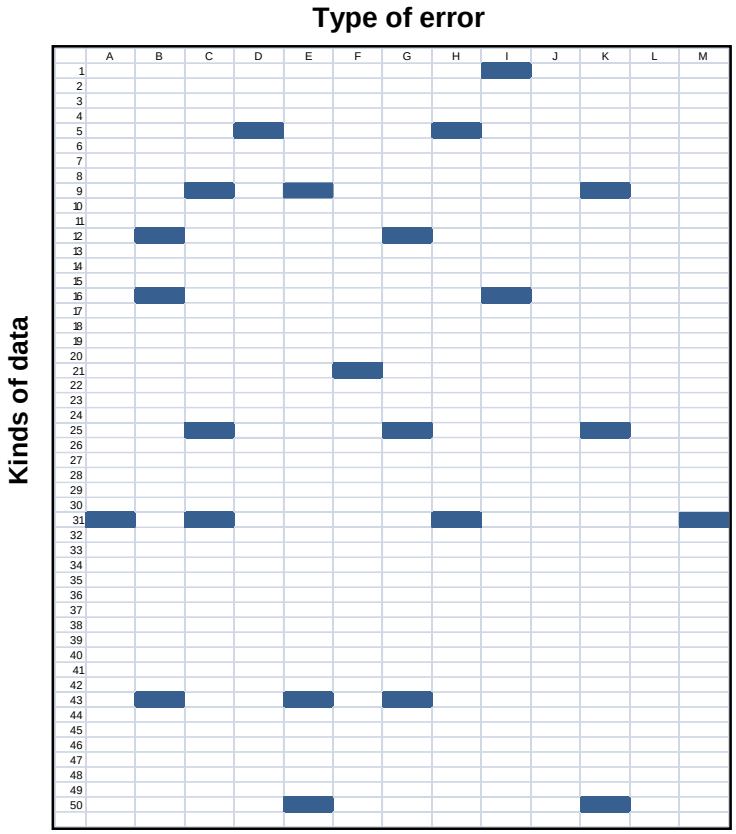
+ Strategy for Improvement

- Strategy for improvement (short term)
  - Define metrics and its relationship to business impact to identify which data to improve
  - Produce baseline
  - Use same metric (periodically/real time) to measure change from baseline
  - Sustain improvements through ongoing monitoring

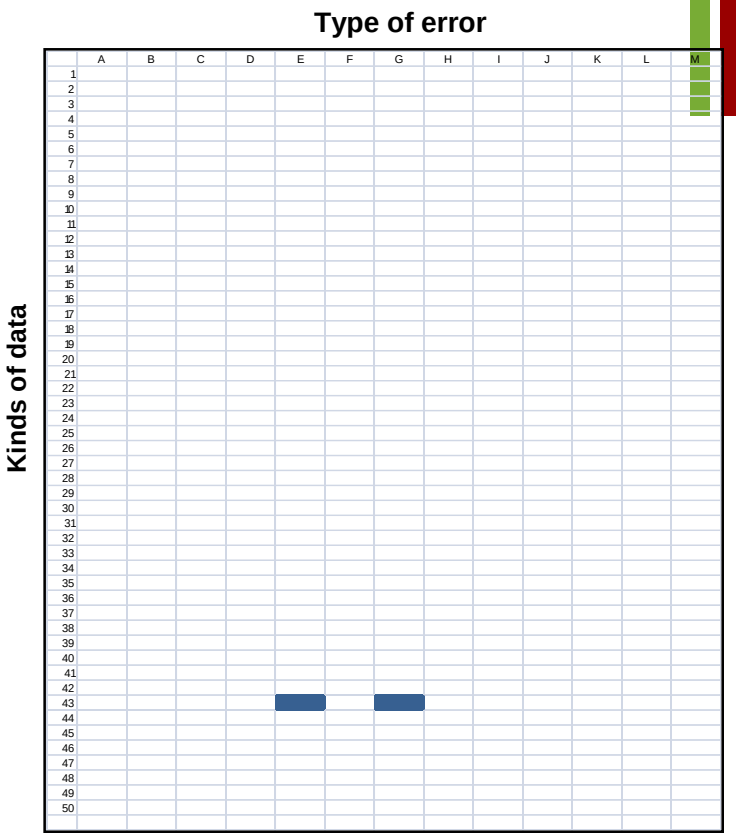
Data Oriented

Harte-hanks Trillium Software (2007) Building a tangible ROI for Data Quality

+ Comparing with Baseline



Before



After

epmo (errors per million opportunities): 3.54% * 1,000,000 = 354,000 *epmo*

epmo (errors per million opportunities): 0.31% * 1,000,000 = 3,100 *epmo*

https://www.sixsigmadaily.com/dpu-dpmo-ppm-and-rty/#:~:text=DPU%20equals%202.Defects%20Per%20Million%20Opportunities%20(DPMO),have%20a%20flaw%20or%20mistake.

+ Strategy for Improvement

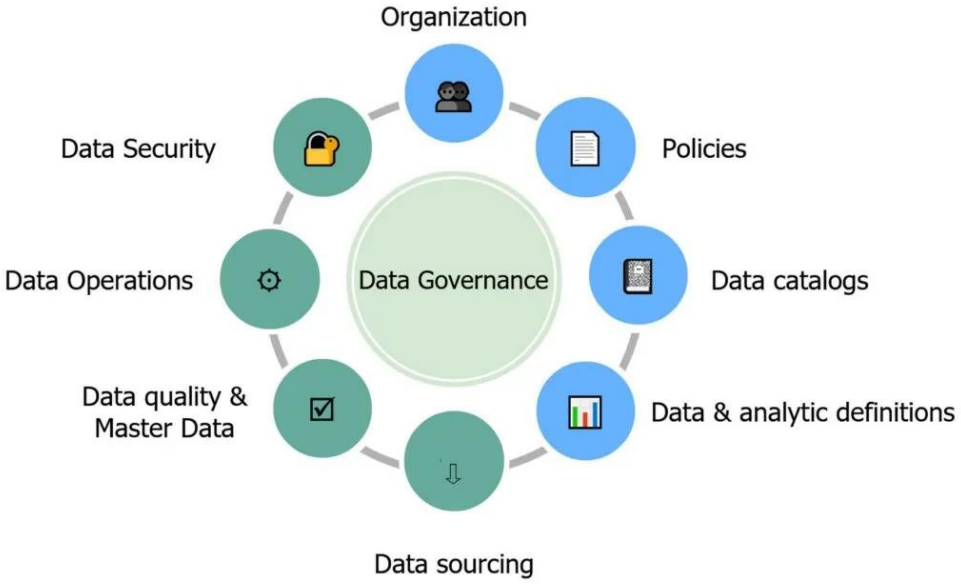
- Strategy for improvement (long term)
  - Establish process owner and management team
  - Describe process qualitatively and understand requirements
  - Establish measurement system
  - Establish process control and conformance to requirements
  - Identify improvement opportunities
  - Select opportunities and set objectives regarding each
  - Make and sustain improvements

Process Oriented

Thomas Redman. Data Quality for the Information Age. 1996

+ Four basic Steps of Data Governance (4-4)

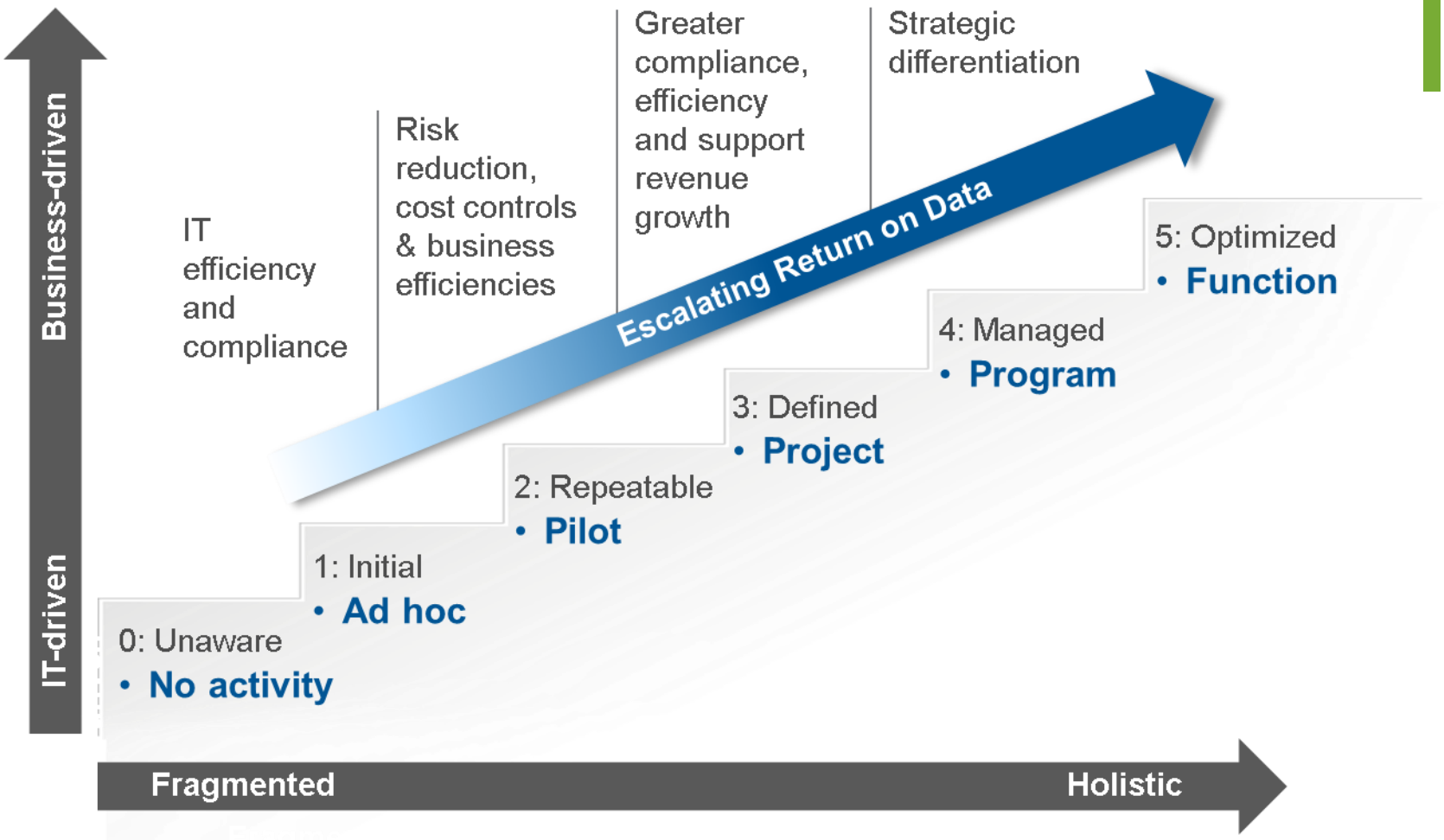
- Recognize the problem
- Measure its costs
- Devise strategy for improvement
- **Aim towards data governance maturity**



https://www.imperva.com/learn/data-security/data-governance/

https://www.avenga.com/magazine/data-governance-what-for-why-and-how/

+ A Data Governance Maturity Model



Many data governance maturity model, from IBM, SAS, DataFlux, Stanford. Google “data governance maturity model and check.

+ Computational Approaches

- Defining and enforcing **data constraints**
  - For describing and enforcing data integrity rules
  - Functional dependency (FD), conditional functional dependency (CFD) and probabilistic CFP
- Doing **entity resolution** effectively and efficiently
  - Linking entities based on similarity, for data integration and duplicate removal
- Managing **missing values** and **uncertain data**
  - **Data imputation** (e.g., data interpolation and augmentation)
  - Managing data with probabilistic values (**probabilistic databases**)
- Data provenance (a.k.a. **data lineage**)
  - Maintaining data origin and processing steps

+ Summary

- Data quality is managed as a product
- Data quality involves many issues, such as *integrity, accuracy, completeness, currency, privacy, availability*, etc.
- Data quality management includes organizational, architectural, and computational solutions
- A key aspect of computations solutions is entity resolution
- There are four steps in data quality management:
  - (1) Recognize the problem
  - (2) Measure its costs
  - (3) Devise strategy for improvement
  - (4) Aim towards data governance maturity

+ Readings

- Wang & Strong, “Beyond Accuracy: What Data Quality means to Data Consumers”, *Journal of MIS* 1996
- Ken Orr, “Data Quality and Systems”, *CACM* 1998
- Guidelines for Setting Organizational Policies for Data Quality, 2008
- Handbook of Data Quality Research and Practice Editors: Sadiq, Shazia (Ed.), 2013

Next Week: Data Security and Privacy

https://www.springer.com/gp/book/9783642362569