

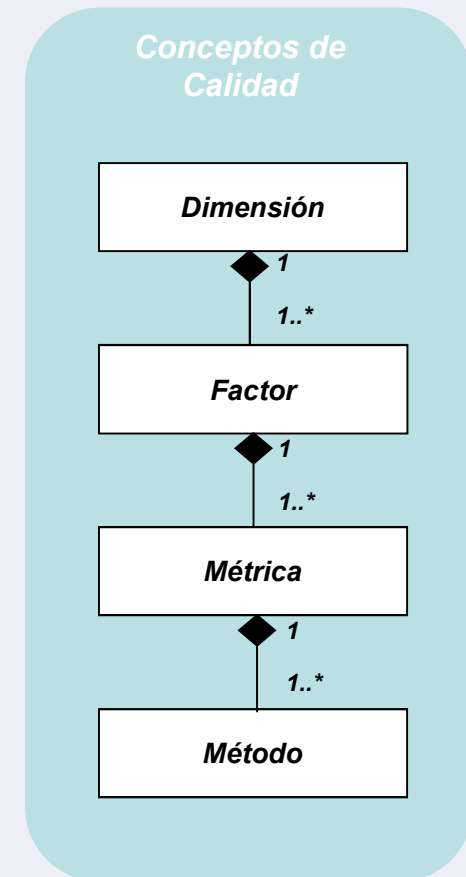
---

# Calidad de Datos e Información

## Métricas de Calidad

# Jerarquía de conceptos de calidad

- Las dimensiones representan las facetas de la calidad a alto nivel.
- Cada dimensión puede refinarse en un conjunto de factores que representan aspectos particulares.
- Cada factor puede medirse con varias métricas.
- Cada métrica puede implementarse con varios métodos de medición.



# Medición de la calidad

- **Métrica de calidad:**
  - Define la forma de medir un factor de calidad
  - Se define con:
    - Un *nombre*
    - Una *descripción* (qué se mide)
      - Ej.: cantidad de valores nulos, cantidad de tuplas, tiempo transcurrido desde la última actualización
    - Las *unidades* de medición
      - Ej.: *tiempo* de respuesta en ms, volumen en GB, un valor entre 0 y 1, etc.
    - La *granularidad* de la medida
      - Fuertemente dependiente del modelo de datos
      - Modelo relacional: celda, tupla, columna, tabla, grupo de tablas, base de datos

## **Modelo Relacional**

Tipo	Producto	Cant	PrecioUnit
Lacteos	Leche	5	1
Lacteos	Yogur	7	1.5
Bebidas	Agua Min	9	0.8

# Métrica de Calidad

---

- Ejemplo Métrica 1:

<b>ExacSemantica-Bool</b>	
Descripción:	Mide si un dato existe en la realidad.
Unidades:	{0,1}
Granularidad:	Celda

- Ejemplo Métrica 2:

<b>Densidad-Grado</b>	
Descripción:	Mide el grado de densidad de una columna.
Unidades:	[0..1]
Granularidad:	Columna

# Medición de la calidad

---

- **Métrica Instanciada:**

- Define la aplicación de una métrica a un dato o conjunto de datos en particular.
- Se define con:
  - Un *nombre*
  - *Datos*
    - Dato o conjunto de datos a los que se aplica la métrica

# Medición de la calidad

- **Método de medición:**

- Un método es un proceso que implementa una métrica.
- Es el encargado de tomar una serie de medidas (correspondientes a una métrica) para una BD concreta.
- La implementación del método depende:
  - de la aplicación en concreto y
  - de la estructura de la BD
- Ejemplo: para medir el tiempo transcurrido desde la última actualización, se puede:
  - Usar timestamps de la BD
  - Acceder a los logs de actualización
  - Comparar versiones de la BD
  - ...

Varios métodos  
de medición



**Una misma métrica puede ser medida por diferentes métodos.**

# Métrica de Calidad

---

- Ejemplo Métrica 1:

<b>ExacSemantica-Bool</b>	
Descripción:	Mide si un dato existe en la realidad.
Unidades:	{0,1}
Granularidad:	Celda

- Para una misma métrica, *ExacSemantica-Bool*, proponemos:
  - 2 métodos (MET1 y MET2), para medir la exactitud semántica del atributo «dirección».

# Métrica Instanciada

- Ejemplo Métrica 1 Instanciada con MET1:

<b>ExacSemantica-Bool-Direccion</b>	
Métrica:	ExacSemantica-Bool
Datos:	Clientes.Direccion
Método Medición:	MET1 - Consultar Google Maps

- Ejemplo Métrica 1 Instanciada con MET2:

<b>ExacSemantica-Bool-Direccion</b>	
Métrica:	ExacSemantica-Bool
Datos:	Clientes.Direccion
Método Medición:	MET2 - Consultar guía Antel



# Métrica de Calidad y Métrica Instanciada

- Ejemplo Métrica 2:

<b>Densidad-Grado</b>	
Descripción:	Mide el grado de densidad de una columna.
Unidades:	[0..1]
Granularidad:	Columna

- Ejemplo Métrica 2 Instanciada:

<b>Densidad-Grado-Ciudad</b>	
Métrica:	Densidad-Grado
Datos:	Cientes.Ciudad
Método Medición:	Contar cantidad de valores NULL en la columna

# Métrica Instanciada

- Ejemplo Métrica 1 Instanciada:

<b>ExacSemantica-Bool-Ciudad</b>	
Métrica:	ExacSemantica-Bool
Datos:	Clientes.Ciudad
Método Medición:	Consultar Google Maps

- Ejemplo Métrica 2 Instanciada:

<b>Densidad-Grado-Ciudad</b>	
Métrica:	Densidad-Grado
Datos:	Clientes.Ciudad
Método Medición:	Contar cantidad de valores NULL en la columna

# Aggregación de medidas

---

- **Medida de calidad**

- Es el valor obtenido por la aplicación de una métrica de calidad durante una medición (Ej., 0 o 1)

- **Granularidad**

- En Modelo Relacional puede ser: Celda, Tupla, Atributo (columna), Tabla (o relación)
- Análogamente en otros modelos
  - En XML/JSON podría ser: cada dato de determinados tags, o un documento.

- **Agregar**

- Obtener a partir de la medida a determinado nivel de granularidad, una medida a un menor nivel de granularidad.
- Ejemplo:
  - A partir de las medidas de las celdas, una medida para toda la tabla.

# Aggregación de medidas

- Ejemplo: dimensión Exactitud, de celda a columna

<u>CI</u>	Nombre	Dirección	Edad
1.234.567-1	María López	Yi 1234	18
2564987-3	René Millán	18 de Julio 322	58
2.345.678-2	José Pérez	Lima 5678	23

# Estudio de algunas dimensiones

---

- Exactitud
- Completitud
- Frescura
- Consistencia
- Unicidad

# Estudio de algunas dimensiones

---


- **Exactitud**
- Completitud
- Frescura
- Consistencia
- Unicidad

# Exactitud (accuracy)

---

- Factores
  - Exactitud semántica (semantic accuracy)
  - Exactitud sintáctica (syntactic accuracy)
  - Precisión (precision)

# Métricas de exactitud

- Tres familias de métricas:
  - Booleanos:
    - Indican si un dato es correcto o no. Valores {0, 1}
    - Ejemplo: un teléfono es válido o no; no hay matices.
  - Funciones de comparación de valores:
    - Miden la cercanía entre un valor  $v$  del SI y un valor  $v'$  correcto.
    - Se pueden normalizar, ej. *distancia* (18153, 18532) / 18532
    - Ejemplo: *edit distance*
  - Grados:
    - Miden el grado de confianza en la exactitud del dato.
    - En gral. se asignan valores entre 0 y 1.
    - Pueden provenir de procesos automáticos de medición o reconocimiento.
    - Ejemplo:  son reconocidos como 'C' con exactitudes 0.80, 1.00 y 0.65 respectivamente.



# Métricas y mediciones de exactitud

---

- La **exactitud semántica** involucra una comparación del SI con el mundo real. Suele ser muy costoso.
  - Ejemplos:
    - Contratar personal que llame por teléfono a todos los clientes y verifique datos.
    - Enviar cartas/emails con promociones que incentiven a los clientes a enviar sus datos.
    - Crear leyes/procedimientos que obliguen a las empresas a declararse
    - Crear procedimientos que obliguen a los estudiantes a registrarse a un curso dejando info de contacto.
- Alternativa:
  - Comparar contra un referencial considerado como válido u otra BD.
  - Ejemplos:
    - Verificar RUTs de empresas contra listado de la DGI
    - Verificar teléfonos de los clientes contra una guía telefónica
    - Verificar datos de empleados contra la BD contable de la empresa

# Métricas y mediciones de exactitud

- Para el factor **Exactitud semántica**

Resumen:

- **Booleano de exactitud semántica:** Si un dato es exacto o no.
- **Desviación de exactitud semántica:** Distancia a los datos exactos.
- **Grado de exactitud semántica:** Impresión de la exactitud semántica de los datos.

*Comparación con la realidad o referencial*

*Asignada por un agente o por un experto*

# Métricas y medición de exactitud

---

- La **exactitud sintáctica** implica verificar si un dato está bien escrito.
- Varias formas de verificar:
  - Por extensión:

Comparar con un **diccionario** que representa el dominio.

    - Ej: los nombres de las calles deben estar en la guía de calles.
    - Algunos dominios son difíciles de representar, ej. apellidos válidos
  - Por comprensión:

Chequear si satisfacen **reglas de sintaxis**.

    - Ejemplos de reglas:
      - Los teléfonos internos tienen 4 dígitos.
      - Las CI deben de verificar el dígito de control.
      - El sexo se almacena como F o M.
      - Las direcciones tienen la forma *calle número apto CP ciudad*.
    - Las reglas aseguran que los datos se almacenan en un mismo formato y son compatibles con datos anteriores.

# Métricas de exactitud

---

- Para el factor **Exactitud sintáctica**

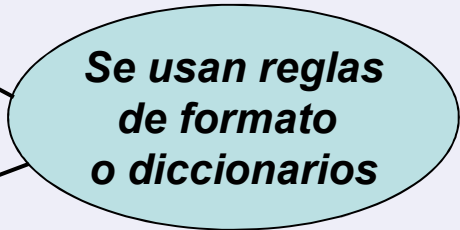
Resumen:

- **Booleano de exactitud sintáctica:**

Si un dato es sintácticamente correcto o no.

- **Desviación de exactitud**

**sintáctica** : Distancia a los datos correctos más parecidos.



*Se usan reglas de formato o diccionarios*

# Ejemplo

---

- Desviación de exactitud sintáctica
  - Función de comparación: *edit distance* (*Levenshtein distance*)
    - Evalúa la distancia del valor almacenado al valor válido más cercano (del dominio correspondiente).
    - Es la mínima cantidad de operaciones (inserciones, borrados y reemplazos de caracteres), necesarios para pasar de un string a otro.

# Métricas y medición de exactitud

---

- **Referenciales:**
  - Usados para verificar la **correctitud semántica.**
  - Contienen un conjunto de parejas <clave, valor>
    - Clave representa un elemento u estado del mundo real
    - Valor representa un atributo de dicho elemento
    - Ej. <CI, nombre>
  - Dos tipos de chequeos:
    - Verificar que la clave pertenezca al referencial (se detectan mismembers).
    - Verificar que la clave esté asociada al valor correcto.
- vs.
- **Diccionarios:**
  - Usados para verificar la **correctitud sintáctica.**
  - Contienen una lista de valores válidos para un dominio.
    - Ej. Nombres de calles
  - El chequeo consiste en verificar que un dato pertenezca al diccionario.

# Métricas de exactitud

- Para el factor **Precisión**:

- **Escala**: Escala de la medición

- Ej.  $87 \pm 1 \text{ cm} \rightarrow 1/87$
    - Ej. “Rojo” vs. “204R-51G-0B”
    - Ej. “Interior” vs. “Colonia”

- **Error estándar**: Desviación estándar de un conjunto de mediciones.

- Ej. medidas de tráfico tomadas por varios sensores

- **Granularidad**: Cantidad y cobertura de los atributos que representan un concepto.

- Ej. calle, número de puerta, ciudad, código postal y país para representar una dirección.

*Precisión del agente o instrumento de medición*

*Jerarquía de precisiones de valores del dominio*

*Varios valores para un mismo estado/entidad*

*Esquema de datos*

*Descomposición de texto libre*

# Ejercicio 6

---

- Considerando los ejercicios realizados anteriormente en clase (Ej. 1 al 5), para algunos de los factores de Exactitud identificados:
  - 1) Dar una **métrica de calidad** de **Exactitud**.
  - 2) **Instanciar** dicha métrica de calidad para algún dato o para algunos datos en particular.



# Agregaciones para exactitud

- Para medir la exactitud de conjuntos de datos (ej. tablas), en función de la exactitud de cada dato se usan:
  - **Ratios:**
    - $\text{AccuracyRatio}(S) = |\{a_i / a_i = 1\}| / n$ 
      - para booleanos
    - $\text{AccuracyRatio}(S) = |\{a_i / a_i \geq \theta\}| / n, 0 \leq \theta \leq 1$ 
      - para distancias y grados
  - **Promedios:**
    - $\text{AccuracyAvg}(S) = (\sum_i a_i) / n$
  - **Promedios con sensibilidades:**
    - $\text{AccuracySens}(S) = (\sum_i a_i^\alpha) / n, 0 \leq \alpha \leq 1$
  - **Promedios ponderados:**
    - $\text{Accuracyweight}(S) = \sum_i w_i a_i, 0 \leq w_i \leq 1, \sum_i w_i = 1$

**Nota:** 1 es un valor correcto, 0 es un valor incorrecto

# Agregaciones para exactitud

- Otras agregaciones posibles:
  - T: tabla que queremos medir, con K atributos y N tuplas.
  - T': una tabla de referencia, donde tenemos el atributo que identifica las tuplas de T, con todos los valores correctos (por ej., ci)
  - $q_{ij}=0$  si la celda ij es correcta,  $q_{ij}=1$  si la celda ij es incorrecta (con  $i:1..N, j:1..K$ )
  - $q_i = \sum_j q_{ij}$ , da la correctitud de la tupla  $t_i$ .
  - $s_i=0$  si la tupla  $t_i$  se corresponde con una tupla en T', si no:  $s_i=1$ .

*weak accuracy error (wae):*  $\sum_i ( \beta((q_i > 0) \wedge (s_i = 0)) / N )$

*strong accuracy error (sae):*  $\sum_i ( \beta(s_i = 1) / N )$

*accuracy (acc):*  $\sum_i ( \beta((q_i = 0) \wedge (s_i = 0)) / N )$

$\beta=1$  si la condición es verdadera,  $\beta=0$  si es falsa

# Ejercicio 7

---

- Dadas las agregaciones de la diapositiva anterior, para c/u:
  - Describa la semántica de cada agregación
  - Analice de qué granularidad a qué otra granularidad pasa cada agregación

# Estudio de algunas dimensiones

---

- Exactitud
- **Completitud**
- Frescura
- Consistencia
- Unicidad

# Compleitud (completeness)

---

- Factores
  - Cobertura (coverage)
  - Densidad (density)

# Métricas y medición completitud

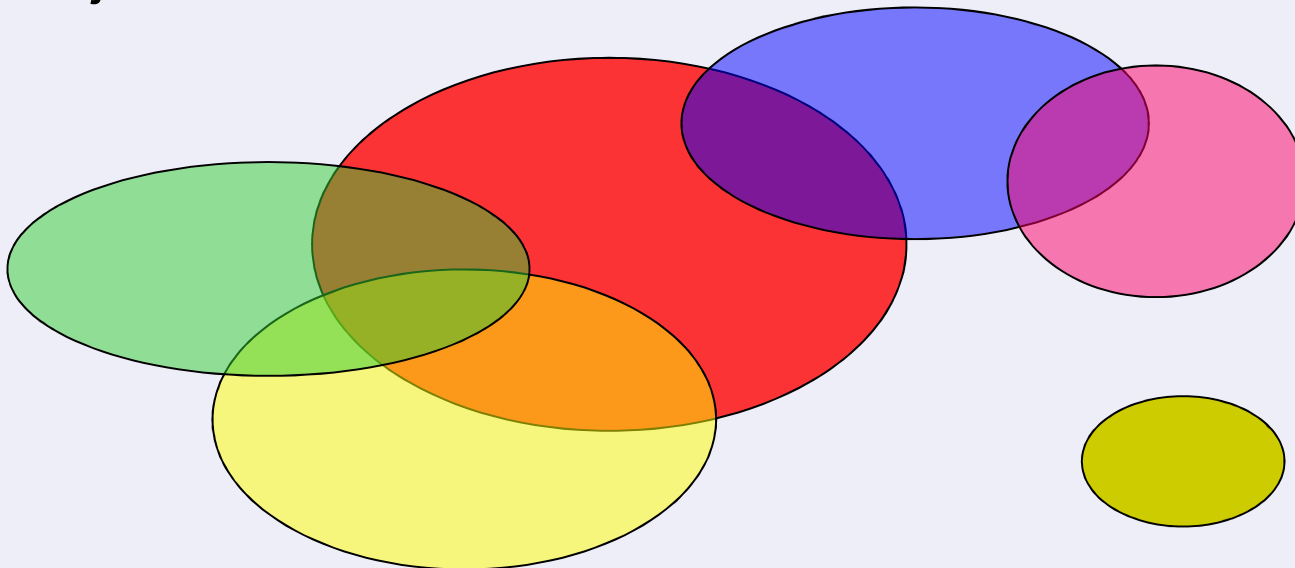
---

- Al igual que la exactitud semántica, la **cobertura** involucra una comparación del SI con el mundo real.
  - Se necesitaría un **referencial**. Pero rara vez es posible obtenerlo.
  - Ejemplos:
    - Identificar todos los clientes potenciales.
    - Conocer todas las personas que están trabajando (legales o ilegales).
    - Tener la lista de todos los habitantes de una ciudad.
- Alternativa: estimar el tamaño que tendría tal referencial.
  - Ejemplos:
    - Clientes potenciales: encuestas de sondeo.
    - Trabajadores: sondeos y estimaciones.
    - Habitantes: datos del último censo.

# Métricas y medición completitud

---

- Estimación del referencial: considerar que la realidad es la unión de los SI conocidos
  - Ejemplos:
    - Todos los vuelos a Paris son los propuestos por las agencias de viajes de Montevideo.



# Métricas de completitud

- Para el factor **Cobertura**:
  - **Ratio de cobertura**: Porcentaje de datos contenidos en el SI.
- Para el factor **Densidad**:
  - **Ratio de densidad**: Porcentaje de valores no nulos.

*Estimación del Referencial o su tamaño*

*Ponderaciones*

## Variantes:

- Ponderando según la importancia de los atributos.
- Ponderando grupos de atributos.
  - Ej. Si no tengo la dirección pero tengo el teléfono no es tan grave.



# Ejercicios 8

---

- Especificar la métrica de calidad para el factor cobertura, presentado en la diapositiva anterior.
- Especificar la métrica definida instanciada para una tabla cualquiera y un método de medición para la misma.
- Agregar a la métrica instanciada una agregación que le parezca interesante.

# Estudio de algunas dimensiones

---

- Exactitud
- Completitud
- Frescura
- Consistencia
- Unicidad

# Frescura (freshness)

---

- Factores
  - Actualidad (currency)
  - Oportunidad (timeliness)
  - Volatilidad (volatility)

# Métricas de frescura

- Para el factor **Actualidad**:
  - **Actualidad1**: Diferencia de tiempo entre el momento de la consulta y la primera modificación no repercutida en el SI.
  - **Actualidad2**: Relación entre: diferencia de tiempo entre el momento de la consulta y la última actualización y Frecuencia de cambio en la realidad (o en el origen).
  - **Booleano de frescura**: El dato está actualizado o no.
- Para el factor **Oportunidad**:
  - **Oportunidad**: Si es actual y llegó a tiempo para la tarea.
- Para el factor **Volatilidad**:
  - Volatilidad: Frecuencia de cambios.

*¿Cuándo ocurrieron los cambios?*

*Logs de los cambios*

*Comparación con el estado actual*

*Momento de tarea involucrada*

# Ejercicio 9

---

- ¿Cómo normalizaría (a un valor entre 0 y 1) las medidas de Actualidad1 y Actualidad2?
- Tomando una tabla de Clientes cualquiera, especificar:
  - Una métrica de calidad para el factor Actualidad
  - La métrica instanciada para algún dato o algunos datos, en particular

# Métrica Oportunidad

---

- Oportunidad (*timeliness*)

$$\text{Max} \left( 0, 1 - \frac{t1 - t2}{t3 - t2} \right)$$

t1: instante en que se entregó el dato al usuario

t2: instante en que el usuario solicitó el dato

t3: instante en que el dato deja de ser útil

# Estudio de algunas dimensiones

---

- Exactitud
- Completitud
- Frescura
- **Consistencia**
- Unicidad

# Consistencia (consistency)

---

- Factores
  - Integridad de dominio (domain integrity)
  - Integridad intra-relación (relation integrity)
  - Integridad inter-relación (inter-relation integrity)



# Métricas de consistencia

---

- **Booleano de Consistencia:**
  - Si el dato satisface o no las reglas (de dominio, intra-relación o inter-relación, según el factor).
  - La granularidad podría ser “celda” o “conjunto de celdas”.
- **Ejemplos:**
  - Si  $0 < edad < 120$   
consistencia = 1  
sino consistencia = 0
  - Si  $años-trabajados < 3 \wedge salario-anual > 25000$   
consistencia = 0  
sino consistencia = 1

# Métricas de consistencia

---

- Agregación:
  - **Ratio de integridad**: Porcentaje de datos que satisfacen las reglas (de dominio, intra-relación o inter-relación, según el factor).
  - Como puede haber varias reglas para una misma relación (o grupo de relaciones), en general se construye una suma ponderada de los resultados de medir dichas reglas.

# Ejercicio 10

---

- Considerando una tabla cualquiera, especificar:
  - Dos **métricas de calidad** para el factor **Integridad intra-relación**.
  - Una **métrica instanciada** para un dato o un conjunto de datos particular, para cada una de las métricas de calidad dadas.
  - Una **combinación** de esas métricas de calidad
  - Una **agregación** que le resulte interesante

# Estudio de algunas dimensiones

---

- Exactitud
- Completitud
- Frescura
- Consistencia
- Unicidad

# Unicidad (uniqueness)

---

- Factores
  - No-duplicación (duplication-free)
  - No-contradicción (contradiction-free)

# Métricas de unicidad

---

- Booleano de Unicidad
  - Si el dato está duplicado o no, para factor No-duplicación
  - Si el dato tiene contradicción o no, para factor No-contradicción
  - La granularidad podría ser “celda” o “conjunto de celdas”.
- Agregaciones
  - Para el factor **No-duplicación**:
    - **Ratio de no-duplicados**: Porcentaje de datos que no están duplicados en forma exacta.
  - Para el factor **No-contracción**:
    - **Ratio de no-contradicciones**: Porcentaje de datos que no están duplicados con contradicciones.



*Técnicas de detección de duplicados*

# Ejercicio 11

---

- Para los datos de una tabla cualquiera, ¿cómo mediría el factor No-Contradicción?
- ¿Qué granularidad tendría la métrica?

# Bibliografía

---

- **Data and Information Quality.** Carlo Batini, Monica Scannapieco. Springer. ISBN: 978-3-319-24104-3. 2016.
- **Data Quality for the Information Age.** Thomas C. Redman. 1996 Artech House Inc., ISBN 0-89006-883-6
- **Information Quality: Fundamentals, Techniques and Use.** Felix Naumann, Kai-Uwe Sattler. EDBT Tutorial, Munich, 2006.
- **Data Quality. The Accuracy Dimension.** Jack E. Olson. Morgan Kaufmann Publishers, Elsevier. 2003. ISBN-10 1-55860-891-5
- **Data Warehouse Institute Survey on Data Quality.** W. Eckerson. Proceedings of the Seventh International Conference on Information Quality (ICIQ-02).
- **The TIQM® Quality System for Total Information Quality Management: Business Excellence through Information Excellence.** Larry English. MIT Information Quality Industry Symposium, 2009.
- **Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “AIMQ: a methodology for information quality assessment,”** Information & management, vol. 40, no. 2, pp. 133–146, 2002.



# Bibliografía

---

- **S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, “Overview and Framework for Data and Information Quality Research,” J. Data and Information Quality, vol. 1, no. 1, pp. 2:1–2:22, Jun. 2009.**
- **D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,” Commun. ACM, vol. 40, no. 5, pp. 103–110, May 1997.**
- **R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” Journal of management information systems, pp. 5–33, 1996.**
- **M. Scannapieco and T. Catarci, “Data quality under a computer science perspective,” Archivi & Computer, vol. 2, pp. 1–15, 2002.**
- **B. Otto, K. M. Huner, and H. Osterle, “Identification of Business Oriented Data Quality Metrics,” presented at the ICIQ, 2009, pp. 122–134.**
- **Y. Lee, S. Madnick, R. Wang, F. Wang, H. Zhang. A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data. MIS Quarterly Executive, 2014.**