Calidad de Datos e Información

Introducción

Temario

- Concepto de calidad de datos
- Problemas, consecuencias y causas de la mala calidad
- Necesidades en distintas áreas de aplicación
- Tipos de datos
- Líneas de Investigación
- Roles y modelos de madurez
- Gestión de la calidad en Sistemas de Información

Datos

- **El valor de los datos**
  - Se han convertido en uno de los activos más importantes y valiosos de las organizaciones
    - Materia prima para toma de decisiones operativas y estratégicas
    - Lo que guía a procesos de todo tipo
  - Análisis en Big data, ingeniería de datos, ciencia de datos
  - En el “mundo globalizado”
    - Web como gran base de datos
    - e-government, e-science, e-learning, e-commerce, e-marketing
  - Datos Abiertos
    - Una de las mayores barreras para su utilización es la calidad de los datos disponibles
    - Falta mucho aún en publicación de datos del Estado
    - Los que se publican no son con suficiente detalle, ni con continuidad en el tiempo

Datos

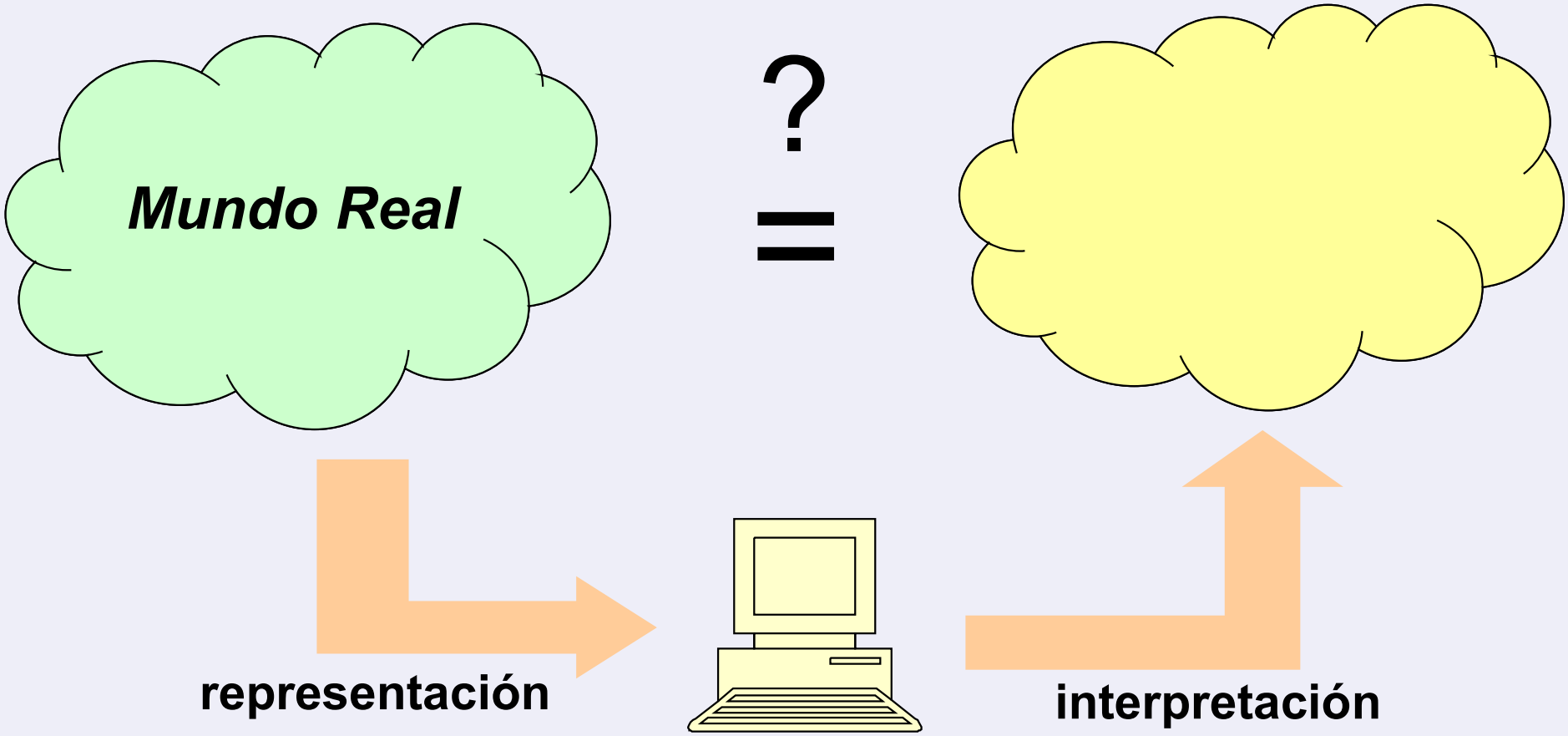
- Distintos tipos de organizaciones los necesitan
  - Su principal actividad es procesar datos
    - compañías de seguros
    - bancos
    - financieras, tarj de credito
    - Etc.
  - Sus actividades y decisiones están guiadas por sistemas de información
    - fábricas
    - proveedores de servicios
    - distribuidores
    - etc.

Datos

- Los datos representan objetos del mundo real en un formato que puede ser
  - almacenado, recuperado y elaborado por un procedimiento de software
  - comunicado a través de una red

- Se pueden clasificar según
  - Su representación
    - estructurado, semi-estructurado, no-estructurado
  - Visión de dato como producto
    - datos brutos (raw data), ítems componentes, información (producto)
  - Complejidad
    - elemental, agregado
  - Otras…

Datos



Calidad

- *"Even though quality cannot be defined, you know what it is"* – Robert Pirsig (filósofo, define la "metafísica de la calidad")

- Calidad de Información – definiciones generales
  - Excelencia / valor
  - Adecuación para su uso (*fitness for use*)
  - Alcanzar o exceder las expectativas del consumidor

- Calidad de Información/Datos es subjetiva
  - Depende del contexto, el consumidor, etc.

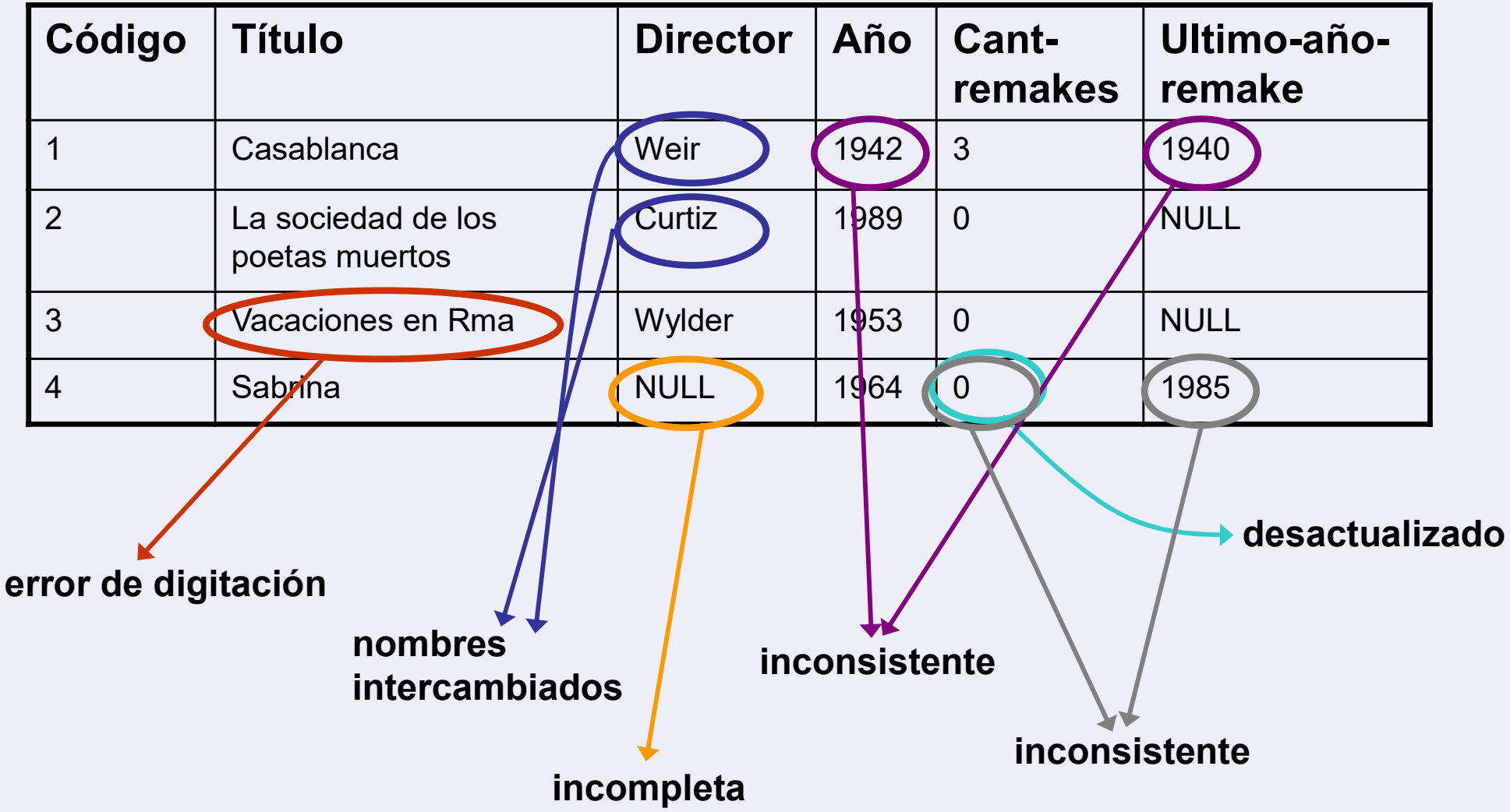
Calidad de Datos

- En general la gente lo reduce a “exactitud de datos” (accuracy), sin embargo…
  - Es un concepto multi-facético, donde existen diferentes dimensiones

- Generalizando, lo que los consumidores quieren de los datos:
  - Que sean relevantes para su uso
  - Que sean correctos y sin inconsistencias
  - Que sean lo más actualizados posible
  - Que se vean en forma adecuada a sus aplicaciones
  - Que se accedan fácilmente

Calidad de Datos - Ejemplo

Código	Título	Director	Año	Cant-remakes	Ultimo-año-remake
1	Casablanca	Weir	1942	3	1940
2	La sociedad de los poetas muertos	Curtiz	1989	0	NULL
3	Vacaciones en Rma	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

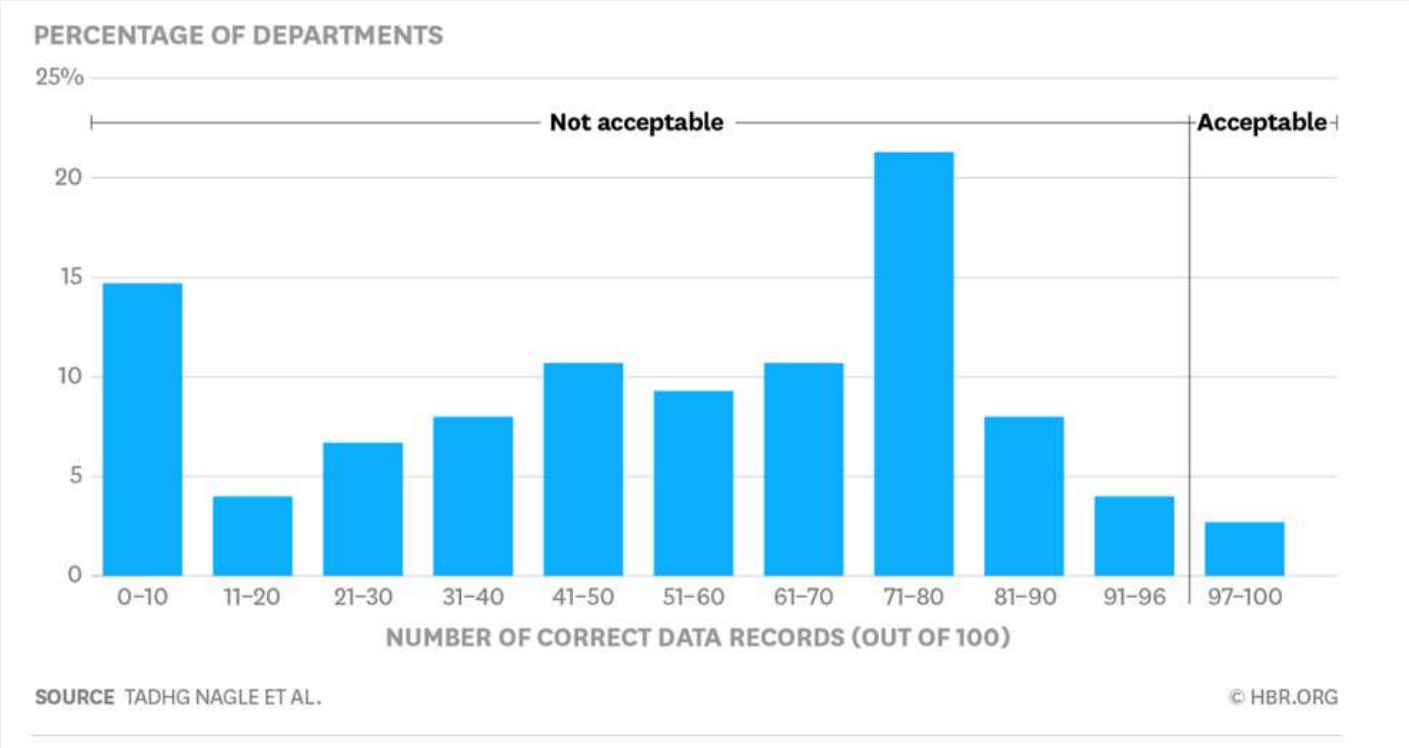


Problemas de calidad

- Datos incorrectos
- Datos inconsistentes con la realidad
- Datos inconsistentes entre sí
- Datos desactualizados
- Información incompleta
- Datos poco confiables debido a su fuente
- Datos difíciles de acceder
- Otros…

Problemas de calidad

- "Only 3% of Companies' Data Meets Basic Quality Standards" [T. Nagle, T. Redman, and D. Sammon 2017]
  - Experiencia en un curso para ejecutivos en Irlanda: ejercicio donde debían elegir 10-15 atributos críticos en los últimos 100 registros de información de su empresa marcando errores obvios.



Problemas de calidad

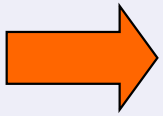
- ¿Qué problemas de calidad de datos de los SI que uds. manejan/mantienen encuentran en su trabajo cotidiano?

- ¿Cómo clasificarían esos problemas según la lista nombrada anteriormente?

Consecuencias de la mala calidad

- Graves consecuencias en la eficiencia y efectividad de organizaciones y negocios

Información con calidad pobre



- ***costos irrecuperables***
- ***reconstrucción de productos y servicios***
- ***soluciones alternativas***
- ***multas***
- ***compensaciones a clientes***

COSTO DIRECTO: llega hasta 15 o 25% del presupuesto de una organización grande

[L. English 2009]

Consecuencias de la mala calidad

- Consecuencias directas - Ejemplos
  - Entregas a clientes en forma tardía o equivocada
  - Errores en el cobro a clientes
  - Clientes duplicados
  - Errores médicos
  - Problemas en implementacion de nuevos sistemas de información que provienen de varias fuentes de datos

Consecuencias de la mala calidad

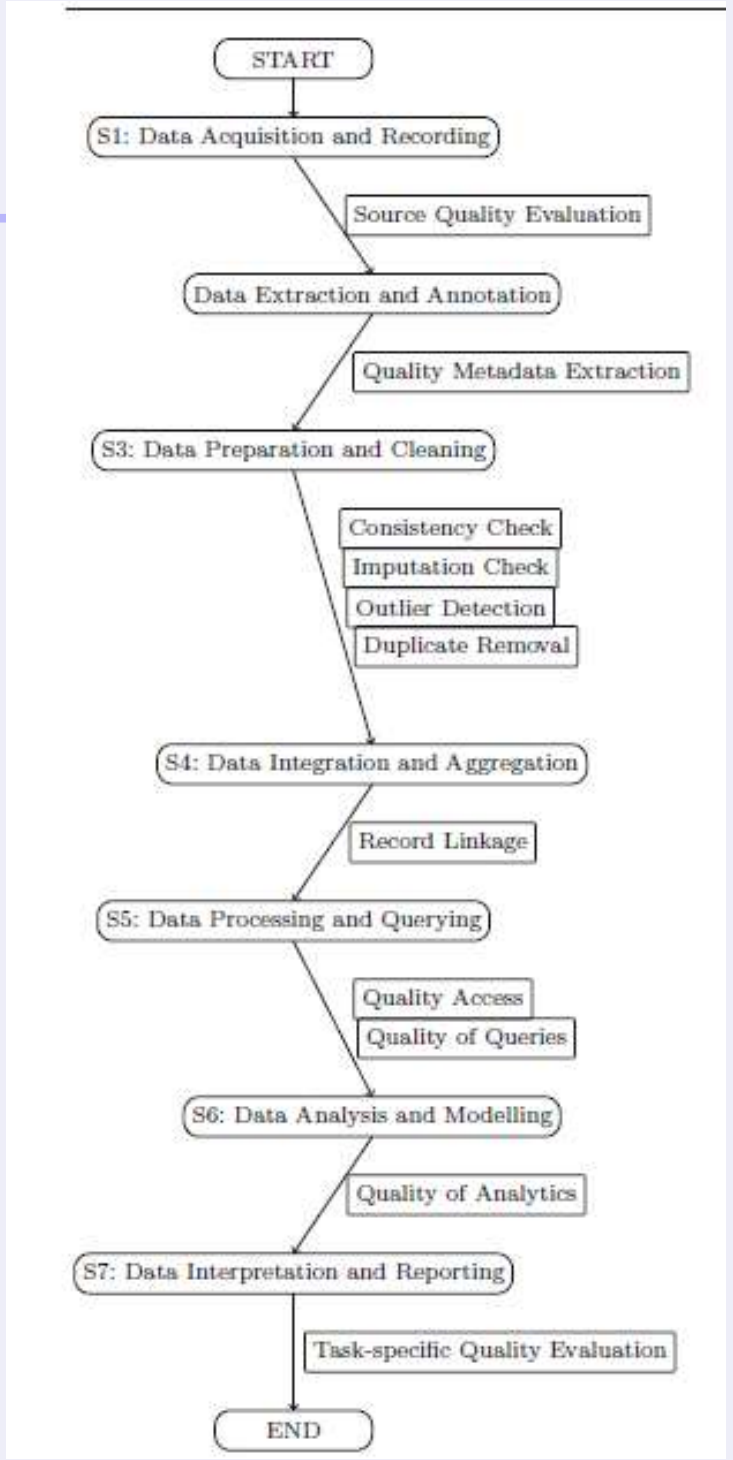
- Consecuencias a mediano/largo plazo - Ejemplos
  - Baja la satisfaccion de los clientes
  - Lleva a costos altos e innecesarios
  - Baja la satisfaccion en el trabajo y alimenta la desconfianza en la organización
  - Tiene impacto en la toma de decisiones
  - Dificulta la re-ingeniería
  - Los avances de la tecnología de la información incrementan el impacto de la mala calidad de datos (disponibilidad a más gente)

Consecuencias en Machine Learning

[Redman 2018]

- Para entrenar un modelo predictivo apropiadamente
  - Los datos históricos deben alcanzar amplios y altos estándares de calidad
  - Los datos deben:
    - Ser correctos (exactos, bien escritos, sin duplicación, etc.)
    - Ser los datos correctos (sin sesgos), cubriendo el rango entero necesario
    - La mayor parte del trabajo en calidad de datos se enfoca en un criterio o en otro , y no en los dos simultánemente.
- Los científicos de datos limpian los datos antes de entrenar el modelo. Tarea tediosa y que insume muchísimo tiempo.
- A medida que las tecnologías de ML penetran en las organizaciones la salida de un modelo predictivo alimentará al siguiente y asi sucesivamente. Un pequeño error de datos en un paso actuará en cascada causando más errores y creciendo a través del proceso entero.

Big Data pipeline



[Ceravolo, … - Big Data Semantics. Journal of Data Semantics. 2018]

Consecuencias de la mala calidad

- Ejemplos conocidos de hace mucho tiempo
  - El “problema del año 2000” tuvo un costo de 1,5 trillones de dólares
  - Desastres de misiones espaciales de la Nasa (*Challenger, Mars Climate Orbiter*) tuvieron como parte importante de las causas, problemas de calidad de datos.
  - Reporte de *DW Institute:* problemas de calidad de datos le cuestan a los negocios del país más de 600 billones de dólares por año. [ICIQ 2002]

Consecuencias de la mala calidad

- ¿Qué consecuencias tienen o podrían tener los problemas de calidad de datos que identificaron anteriormente?

Causas de la mala calidad

- Los problemas podrían generarse durante
  - Producción de los datos
  - Procesamiento
  - Almacenamiento
  - Utilización

Causas de la mala calidad

- ¿Cuáles serían las causas de los problemas identificados anteriormente?

Causas de la mala calidad

- Producción de los datos
  - Recolección de datos mediante ingreso humano
  - Problemas sistemáticos con la recolección de datos
  - Diferentes fuentes con representaciones diferentes del mismo objeto de la realidad
  - No mantenimiento al día de los datos
  - Ausencia de un responsable de los datos y de su calidad
- Procesamiento
  - Transformaciones a otras estructuras y formatos
  - Cálculos con datos de entrada, como resúmenes y cálculos de indicadores
  - Unión de datos provenientes de varias fuentes

Causas de la mala calidad

- Almacenamiento
  - Formatos diferentes
  - Ausencia de formatos definidos
  - Bases de datos mal diseñadas

- Utilización
  - Capacidad de análisis y procesamiento insuficiente
  - Cambios en los requerimientos de calidad
  - Uso equivocado de los datos, por mala interpretación o aplicación fuera de contexto
  - Problemas de seguridad y acceso
  - Mal diseño de los sistemas que procesan los datos para su análisis posterior

Necesidades en áreas de aplicación

- Gobierno electrónico
  - Relación gobierno/agencias-ciudadanos/empresas a través de tecnologías de información y comunicación
  - Datos Abiertos de Gobierno
  - Problema principal
    - información similar sobre un ciudadano o una empresa suele estar en múltiples bases de datos.
  - Errores comunes
    - Datos de ciudadanos no actualizados, ingresados con errores
    - Diferentes formatos en las diferentes fuentes (bds de agencias)
  - Consecuencias negativas
    - Inconsistencias entre registros que corresponden al mismo ciudadano o empresa dificultan
      - El servicio que se da al ciudadano
      - Referencia cruzada entre agencias para detectar fraudes, etc.

Necesidades en áreas de aplicación

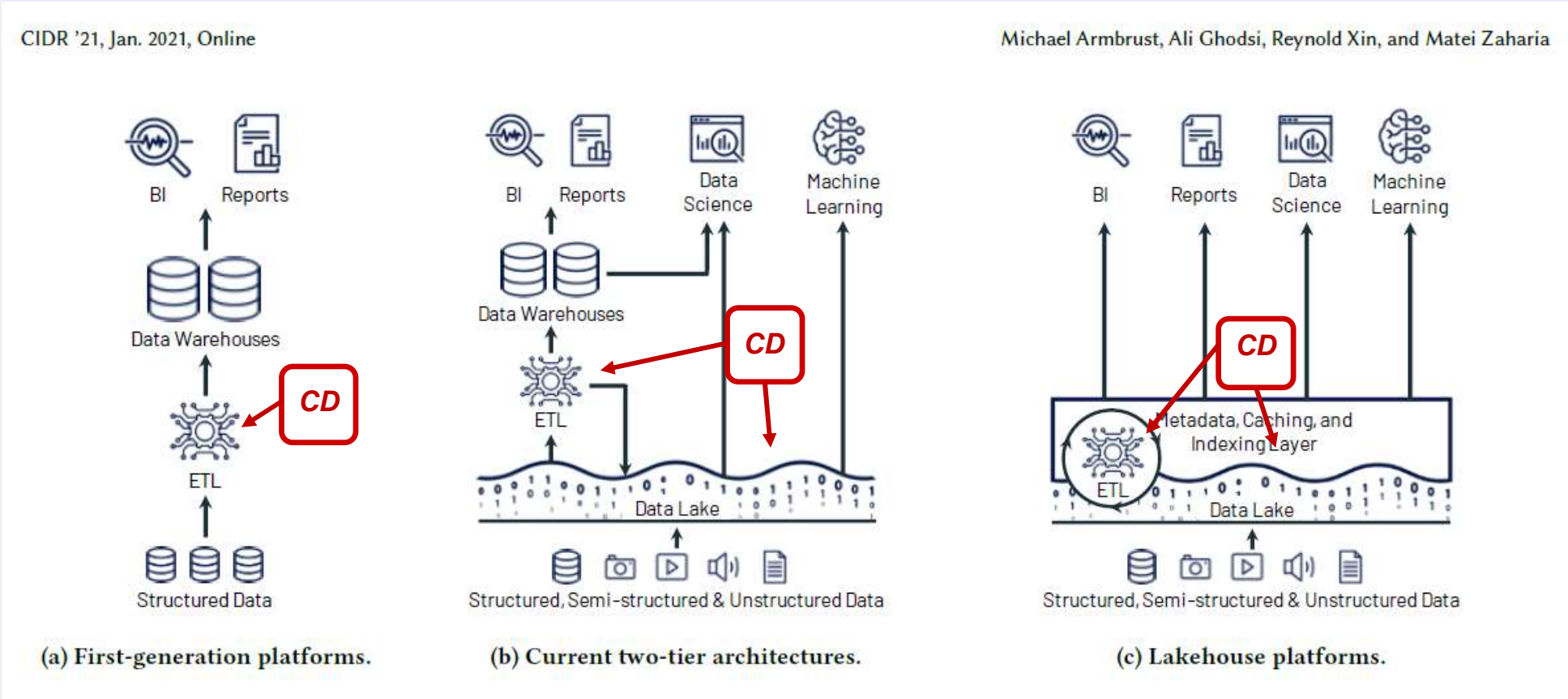
- Bioinformática
  - Se analizan datos genómicos, por ejemplo, para encontrar relaciones con fenotipos o enfermedades específicas
  - Grandes volúmenes de datos
  - Diversos tipos de datos
  - Muchas fuentes heterogéneas y desconectadas
  - Calidad muy variada
  - Los biólogos analizan la calidad manualmente integrando y resolviendo contradicciones entre los datos

Necesidades en áreas de aplicación

- Sistemas de Información proveniente de la Web
  - Presentan gran cantidad de datos
  - Gran cantidad de consumidores
  - Calidad muy heterogénea
  - Variedad de formatos, por ej., csv, xml, linked data
  - Un sitio web es una fuente de datos en continua evolución
  - Es muy difícil individualizar el “dueño de los datos”, responsable de los mismos
  - Caso particular de estos sistemas: Web Warehouse

Necesidades en áreas de aplicación

- Arquitecturas de Big Data



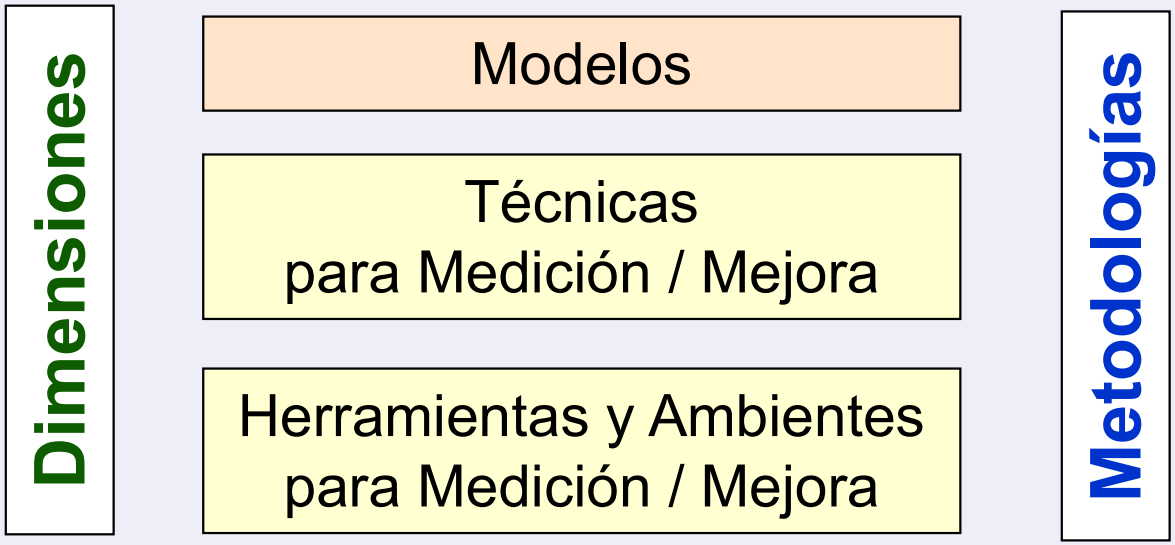
Data Preparation

Tipos de datos en Big Data

- 3 tipos principales de fuentes de big data [Batini 2015]
  - *human sourced*
    - experiencia humana
    - redes sociales, blogs, resultados de búsquedas en internet, videos, mapas, fotos, etc.
  - *process mediated*
    - registros de datos producidos en los procesos de negocio
    - usualmente estructurada y almacenada en bds relacionales
  - *machine generated*
    - datos provenientes de sensores y máquinas utilizados para medir y registrar eventos del mundo físico (IOT)
    - bien estructurados pero de gran tamaño y velocidad

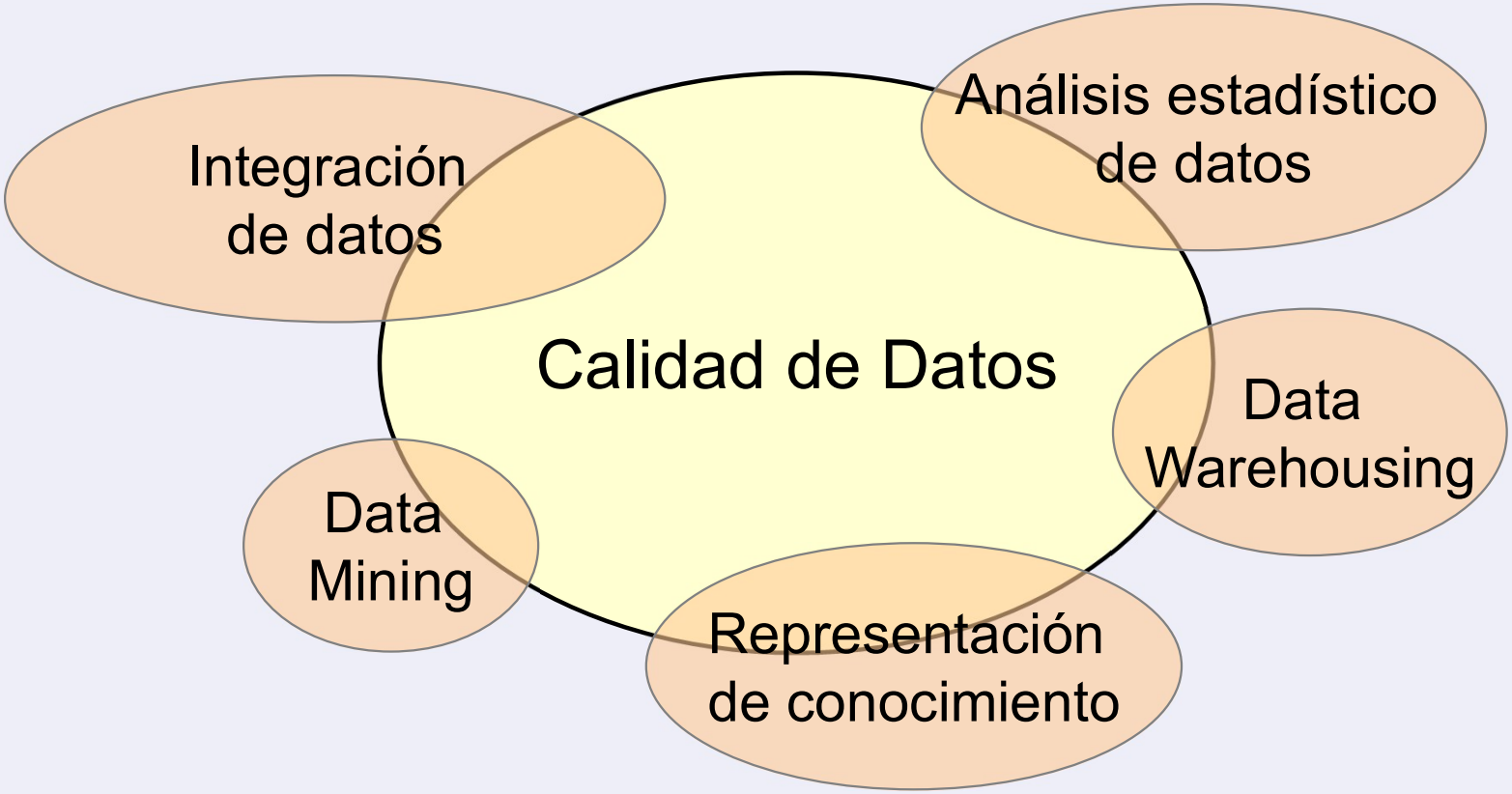
Líneas de Investigación

- Sistema de Gestión de la Calidad
  - Conjunto de técnicas, servicios y herramientas para manejar la calidad de los datos en una organización o varias cooperando.



Líneas de Investigación

- Temas / Áreas de Investigación relacionados con Calidad de Datos



Roles en Calidad de Datos

- **Chief Data Officer (CDO)** [wikipedia]
  - Es el responsable de la governanza y utilización de la información de toda la empresa, como un activo, a través del procesamiento, análisis, minería y comercio de datos.
  - Reporta al CEO
  - Es un gerente ejecutivo
- 90% de las grandes organizaciones tendría un CDO en 2019, según Gartner
  - La carrera por competitividad y eficiencia a través de utilizar la información como un activo está llevando a un crecimiento abrupto de la cantidad de CDOs.
- A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data [Lee 2014]

Roles en Calidad de Datos
[Otto 2009]

- Roles
  - Chief Data Steward
    - Responsable de la calidad de los datos
  - Business Data Steward
    - Pueden haber varios, para los distintos procesos de negocio
  - Technical Data Steward
    - Experto en los SI que manejan los datos. Pueden haber varios.
  - Process Owner
    - Se preocupa por la mejora de los procesos
  - Process User
    - La calidad de datos lo impacta en su trabajo cotidiano.
  - Sponsor
    - Toma las decisiones que luego el *chief data steward* pone en práctica

Modelos de Madurez en CD

- Establecen criterios para definir en qué nivel de madurez con respecto a la Calidad de Datos está la organización.

P6

INFORMATION QUALITY MANAGEMENT MATURITY GRID

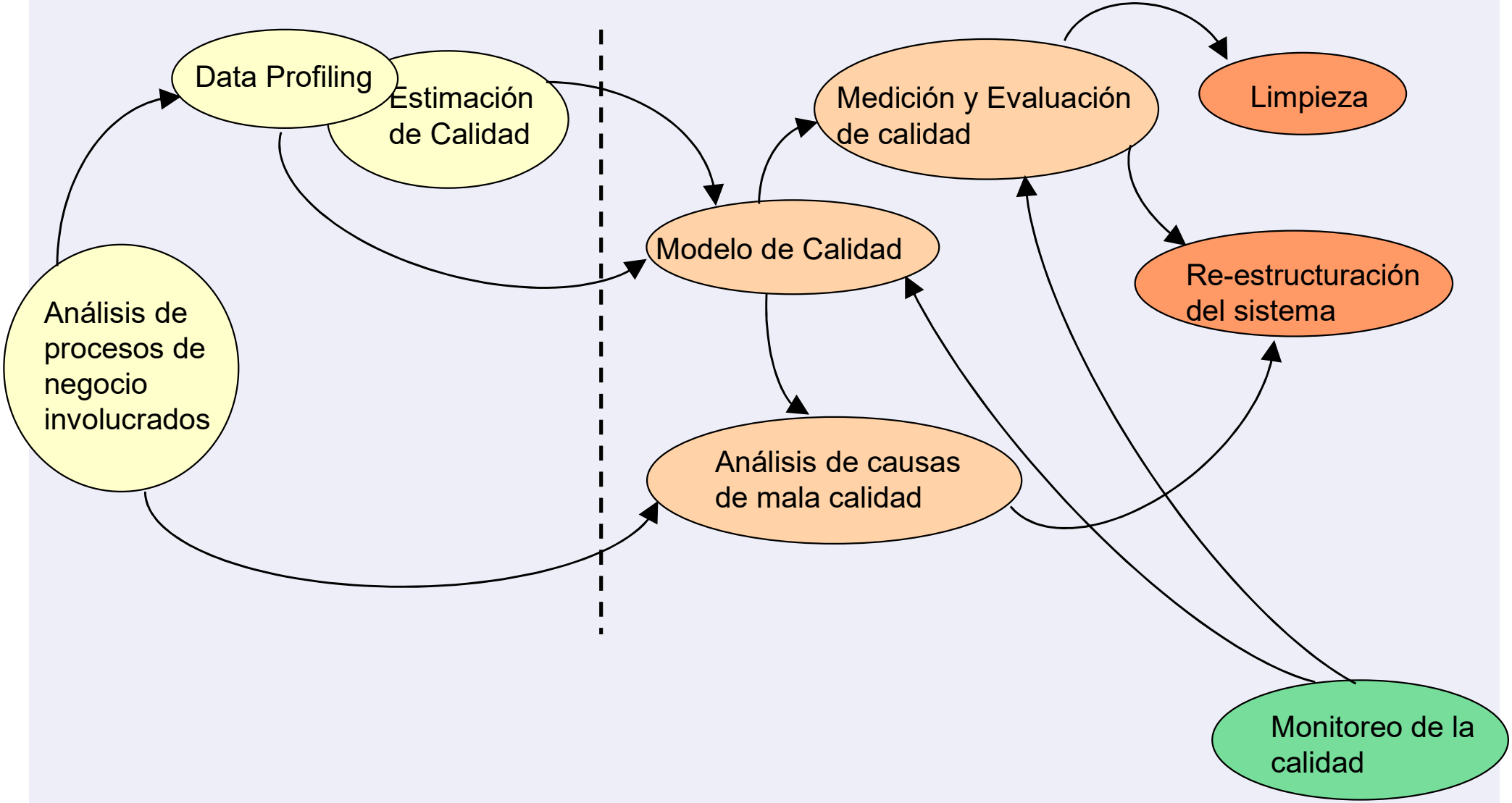
Measurement Categories	Stage 1: Uncertainty (Ad hoc)	Stage 2: Awakening (Repeatable)	Stage 3: Enlightenment (Defined)	Stage 4: Wisdom (Managed)	Stage 5: Certainty (Optimizing)
1. Management understanding and attitude	No comprehension of information quality as a management tool. Tend to blame Data Management or I/S org for "information quality problems" or vice versa.	Recognizing that information quality management may be of value but not willing to provide money or time to make it all happen.	While going through information quality improvement program learn more about quality management; becoming supportive and helpful.	Participating. Understand absolutes of information quality management. Recognize their personal role in continuing emphasis.	Consider information quality management an essential part of company system.
2. Information quality organization status	"Data" quality is hidden in application development departments. Data audits probably not part of organization. Emphasis on correcting bad data.	A stronger information quality role is "appointed" but main emphasis is still on correcting bad data.	Information quality organization exists, all assessment is incorporated and manager has role in development of applications.	Information quality manager reports to CIO; effective status reporting and preventive action. Involved with business areas.	Information quality manager is part of management team. Prevention is main focus. Information quality is a thought leader.
3. Information quality problem handling	Problems are fought as they occur; no resolution; inadequate definition; lots of yelling and accusations.	Teams are set up to attack major problems. Long-range solutions are not solicited.	Corrective action communication established. Problems are faced openly and resolved in orderly way.	Problems are identified early in their development. All functions are open to suggestion & improvement.	Except in the most unusual cases, information quality problems are prevented.
4. Cost of information quality as % of revenue	Reported: unknown Actual: 20%	Reported: 5% Actual: 18%	Reported: 10% Actual: 15%	Reported: 8% Actual: 10%	Reported: 5% Actual: 5%
5. Information quality improvement actions	No organized activities. No understanding of such activities.	Trying obvious "motivational" short-range efforts.	Implementation of the 14 point program with thorough understanding and establishment of each step.	Continuing the 14 point program and starting to optimize.	Information quality improvement is a normal and continued activity.
Summation of company information quality posture	"We don't know why we have problems with information quality."	"Is it absolutely necessary to always have problems with information quality?"	"Through management commitment and information quality improvement we are identifying and resolving our problems."	"Information quality problem prevention is a routine part of our operation."	"We know why we do not have problems with information quality."

Adapted from P. B. Crosby
Quality Management Maturity Model

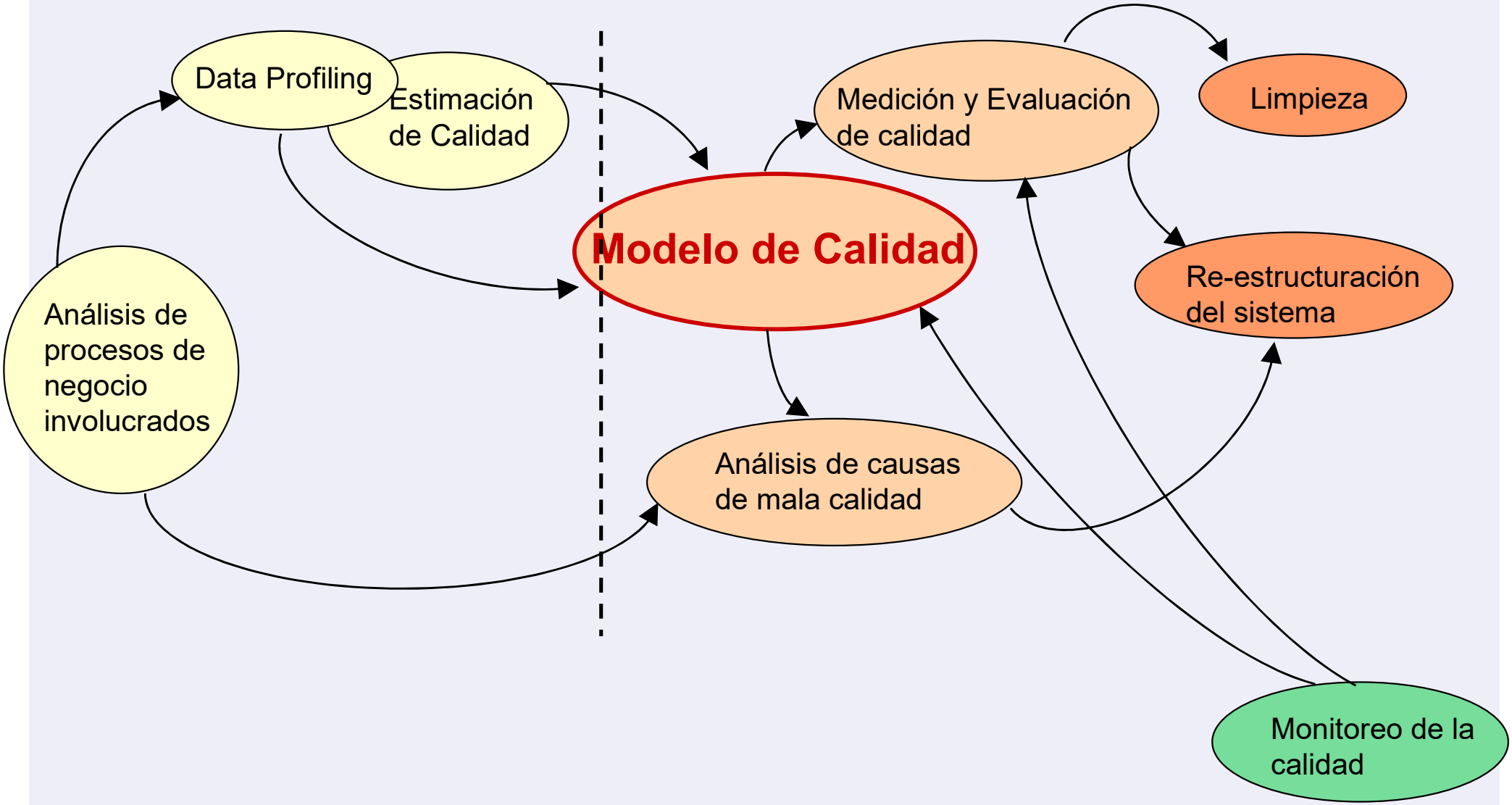
 L. English, *Improving Data Warehouse and Business Information Quality*, pg. 428

Categorías de Medición	Nivel 1: Incertidumbre	Nivel 2: Concientización	Nivel 3: Iluminación	Nivel 4: Conocimiento	Nivel 5: Certeza
1. Comprensión de la gestión de calidad y actitud	No se considera la calidad de la información como una herramienta de gestión.	Se reconoce que la gestión de calidad de la información puede ser de valor, pero no se está dispuesto a invertir dinero y/o tiempo.	Durante el proceso de la mejora de la calidad de la información. La gestión de calidad de la información se vuelve útil y de ayuda.	Se comprende principios de gestión de calidad de información. Se reconoce el rol personal.	Se considera la gestión de calidad de la información una parte esencial de la compañía.
2. Estado de la calidad de la información en la organización	La calidad de los datos está escondida en el desarrollo de las aplicaciones.	Se conoce el rol de la calidad de la información pero el esfuerzo está en corregir los datos de mala calidad.	Existe la calidad de la información en la organización, la evaluación de la calidad está incorporada.	El administrador de la calidad de la información reporta al CIO. La calidad se relaciona con las áreas de negocio	El administrador de la calidad forma parte del equipo admiistrativo. La prevención es el foco principal
3. Manejo de problemas de calidad	Los problemas se resuelven a medida que ocurren.	Los equipos atacan problemas fundamentales. No se solicitan soluciones a largo plazo	Los problemas se enfrentan correctamente, y se resuelven de forma ordenada	Los problemas son identificados a tiempo en el desarrollo.	A excepción de casos inusuales, los problemas de calidad de la información están previstos.
4. Costo de los problemas de calidad como % de los ingresos	Reportado: desconocido Costo: 20%	Reportado: 5% Costo 18%	Reportado: 10% Costo 15%	Reportado:8% Costo: 10%	Reportado: 5% Costo: 5%
5. Acciones de mejora de calidad de la información	No hay actividades organizadas.	Se intenta hacer esfuerzos motivacionales a corto plazo	Implementación de programa de mejora de calidad con entendimiento y establecimiento de cada punto	Se continua el programa mejora de calidad y se comienza a optimizar	La mejora de la calidad de la información es una actividad normal y continua
6.Postura de la compañía	No sabemos por qué tenemos problemas con la calidad de la información	Es necesario siempre tener problemas con la calidad de la información	Mediante gestión, compromiso y mejora de la calidad de la información estamos identificando y resolviendo problemas	La prevención de los problemas de calidad de la información es parte de la rutina de nuestra operación	Sabemos por qué no tenemos problemas con la calidad de la información

Gestión de la calidad en SI



Gestión de la calidad en SI



Modelo de Calidad de datos

- Define:
  - qué características de calidad se manejan
  - sobre qué datos aplican
  - cómo se miden esas características
- Para cada conjunto de datos se define un modelo de calidad particular
- Guía toda la gestión de la calidad de los datos

Bibliografía

- **Carlo Batini, Monica Scannapieco. Data and Information Quality. Springer. ISBN: 978-3-319-24104-3. 2016.**
- **Thomas C. Redman. Data Quality for the Information Age. 1996 Artech House Inc., ISBN 0-89006-883-6**
- **G. Shankaranarayanan y R. Blake. From Content to Context: The Evolution and Growth of Data Quality Research. J. Data and Information Quality, vol. 8, n.º 2, p. 9:1–9:28, 2017.**
- **Jack E. Olson. Data Quality. The Accuracy Dimension. Morgan Kaufmann Publishers, Elsevier. 2003. ISBN-10 1-55860-891-5**
- **W. Eckerson. Data Warehouse Institute Survey on Data Quality. Proceedings of the Seventh International Conference on Information Quality (ICIQ-02).**
- **Larry English. The TIQM® Quality System for Total Information Quality Management: Business Excellence through Information Excellence. 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.**
- **R. M. Pirsig, Zen and the art of motorcycle maintenance : an inquiry into values. HarperPerennial, 2008. ISBN: 978-0-06-167373-3**