Ciencias de Datos

Jose Aguilar

UNIVERSIDAD DE LOS ANDES
MERIDA VENEZUELA

Contenido

¡Todos estamos generando datos de eventos!

tomando el tren

abastecimiento de combustible el automóvil

comprando un café

ajustando la temperatura

poniendome una multa por exceso de velocidad

enviando un correo electrónico

haciendo una cita

viendo esta clase

Mi telefóno inteligente

Más de 14 sensores

Acelerómetro

GPS

WIFI

Bluetooth

Sensor Proximidad

Sensor de luz ambiental

Sensor giroscópico

Pantalla táctil

Cámara(s)

Internet de eventos: 4 fuentes de datos de eventos

¡Las preguntas de análisis de datos están en todas partes!

Necesidad de científicos de datos!

Un científico de datos puede recopilar, analizar e interpretar datos de una variedad de fuentes (interacción social, procesos comerciales, sistemas ciberfísicos).

Convierte datos en valor!!

Cuatro preguntas genéricas sobre ciencia de datos

- ¿Que pasó? — Reportar
- ¿Por qué sucedió? — Diagnósticar
- ¿Que sucederá? — Predicir
- ¿Qué es lo mejor que puede pasar? — Recomendar

Una dimensión local de la Industria 4.0

¿Cómo se usan realmente equipos?

¿Qué componentes deberían ser reemplazados?

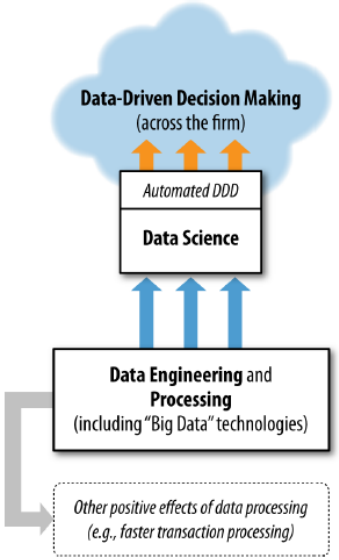
¿Por qué y cuándo funcionan mal equipos?

¿Qué partes necesitan mejorarse?

¿Podemos predecir que equipos se descompondrá la próxima semana?

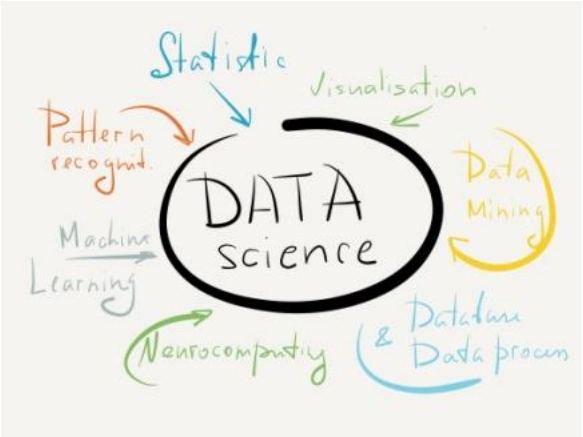
La ciencia de los datos

la ciencia de datos requiere de principios, procesos y técnicas para la comprensión de los fenómenos a través del análisis (automatizado) de los datos.



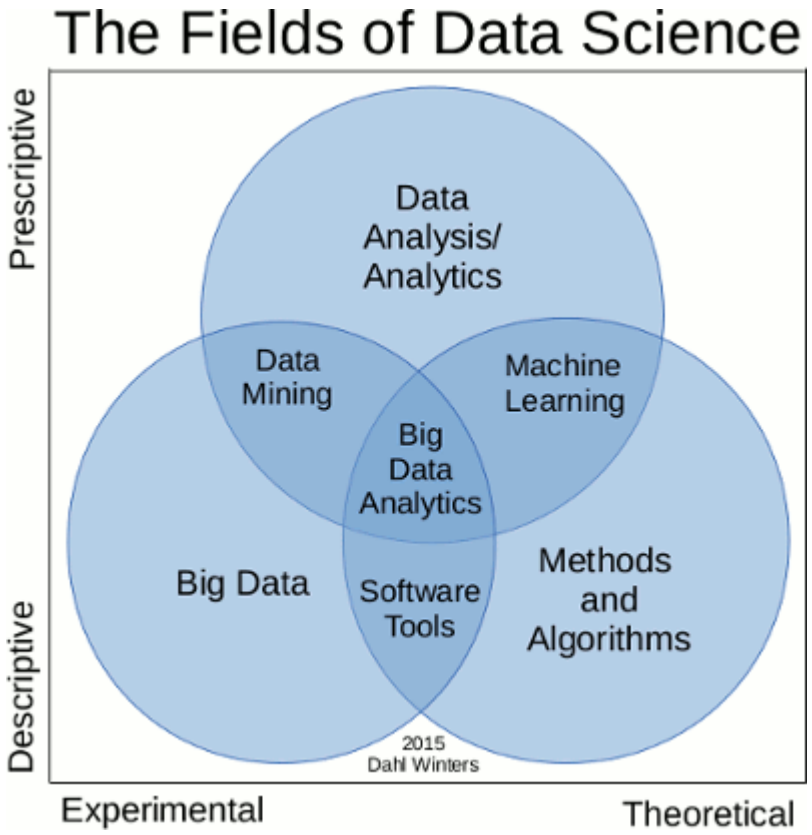
Ciencia de Datos

Ámbito del conocimiento que engloba las habilidades asociadas al análisis inteligente de datos

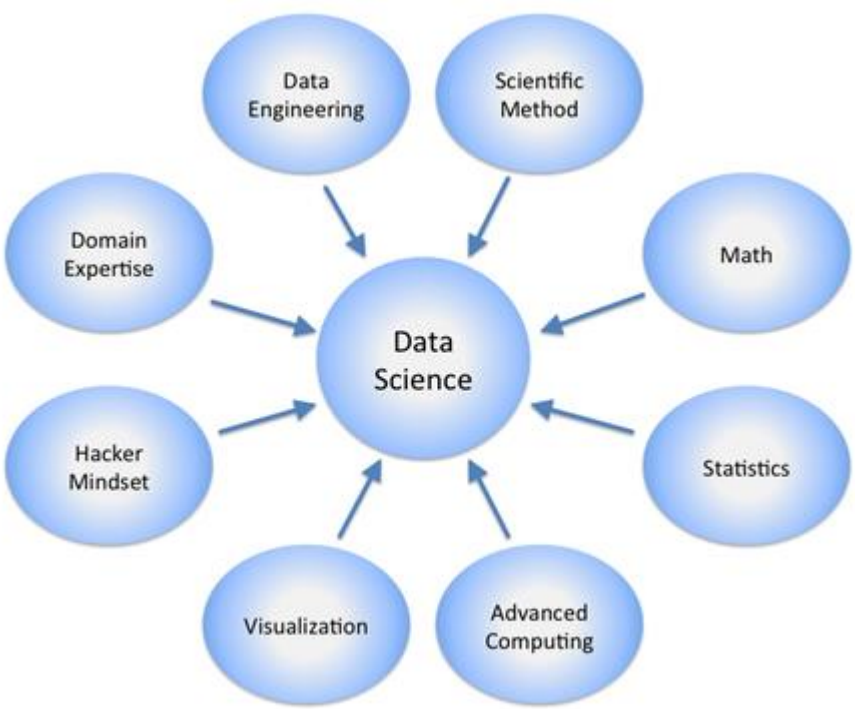


Científico de datos

La ciencia de los datos



La ciencia de los datos



Combinación de las matemáticas, estadísticas, etc., para resolver el problema de captura de datos, además de la limpieza, la preparación y la alineación de los datos.

estadísticas
estocástica

conocimiento del
dominio

Ingeniería
industriales

comportamiento /
social

**minería
datos**

Ciencias de los datos

**minería
proceso**

bases de datos
Algoritmos

inteligencia
proceso
negocio

computación distribuida
a gran escala

visualización
datos

Vista del proceso centrada en la ciencia de datos

Comprender y mejorar el comportamiento y el rendimiento de los equipos utilizando terabytes de datos de eventos de diferentes niveles (nano, micro, macro) recopilados a través de una red de sensores y servicios remotos

Introducción

Etapas de un estudio basado en datos

etapas habituales en el llamado método científico cuyas etapas son:

1) **Planteamiento del problema:** consiste en definir el objetivo de la investigación y precisar el universo o población.

2) **Recogida de la información:** consiste en recolectar los datos necesarios relacionados al problema de investigación.

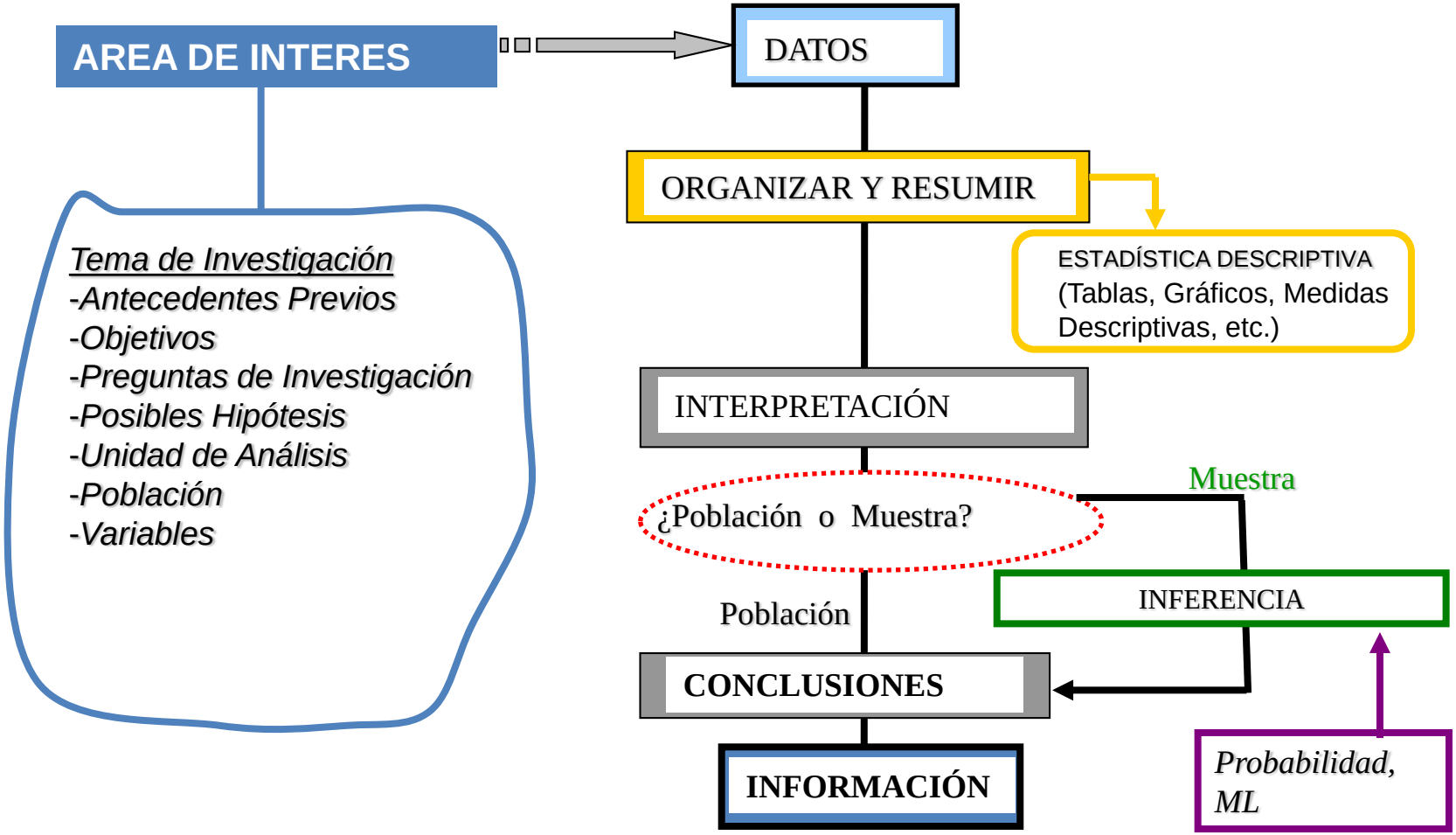
3) **Análisis descriptivo:** consiste en resumir los datos disponibles para extraer la información relevante en el estudio.

4) **Inferencia:** consiste en suponer un modelo para toda la población partiendo de los datos analizados para obtener conclusiones generales.

5) **Diagnóstico:** consiste en verificar la validez de los supuestos del modelo que nos han permitido interpretar los datos y llegar a conclusiones sobre la población

Introducción

Esquema de las etapas de un estudio basado en datos



Introducción

Ejemplos de algunos problemas a estudiar

1) **Se quiere estudiar si en cierto colectivo existe discriminación salarial debida al sexo de la persona empleada.**

2) **Se quiere determinar el perfil de los trabajadores en términos de condiciones económicas y sociales en diferentes comunidades.**

3) **Se quiere estudiar el consumo de las personas de una zona determinada en cuanto a vestuario, alimentación, ocio y vivienda.**

4) **Se quiere determinar las tallas estándar en vestuario para mujeres españolas.**

5) **Se quiere determinar el tiempo que dedican al trabajo y a la familia los trabajadores de distintas empresas del país.**

6) **Se quiere determinar el perfil sociodemográfico de los estudiantes de una Universidad.**

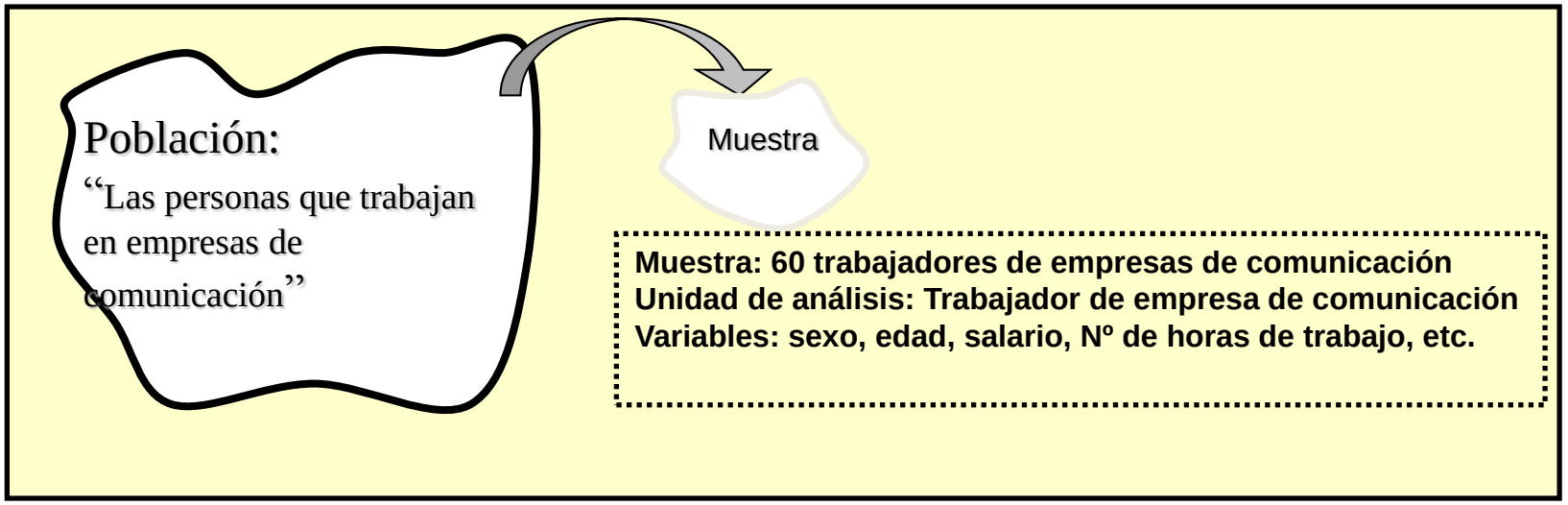
7) **Se quiere estudiar el gasto en teléfono móvil mensual de los estudiantes de una Universidad, y si éste tiene alguna relación con su edad u otras características.**

Algunos conceptos claves

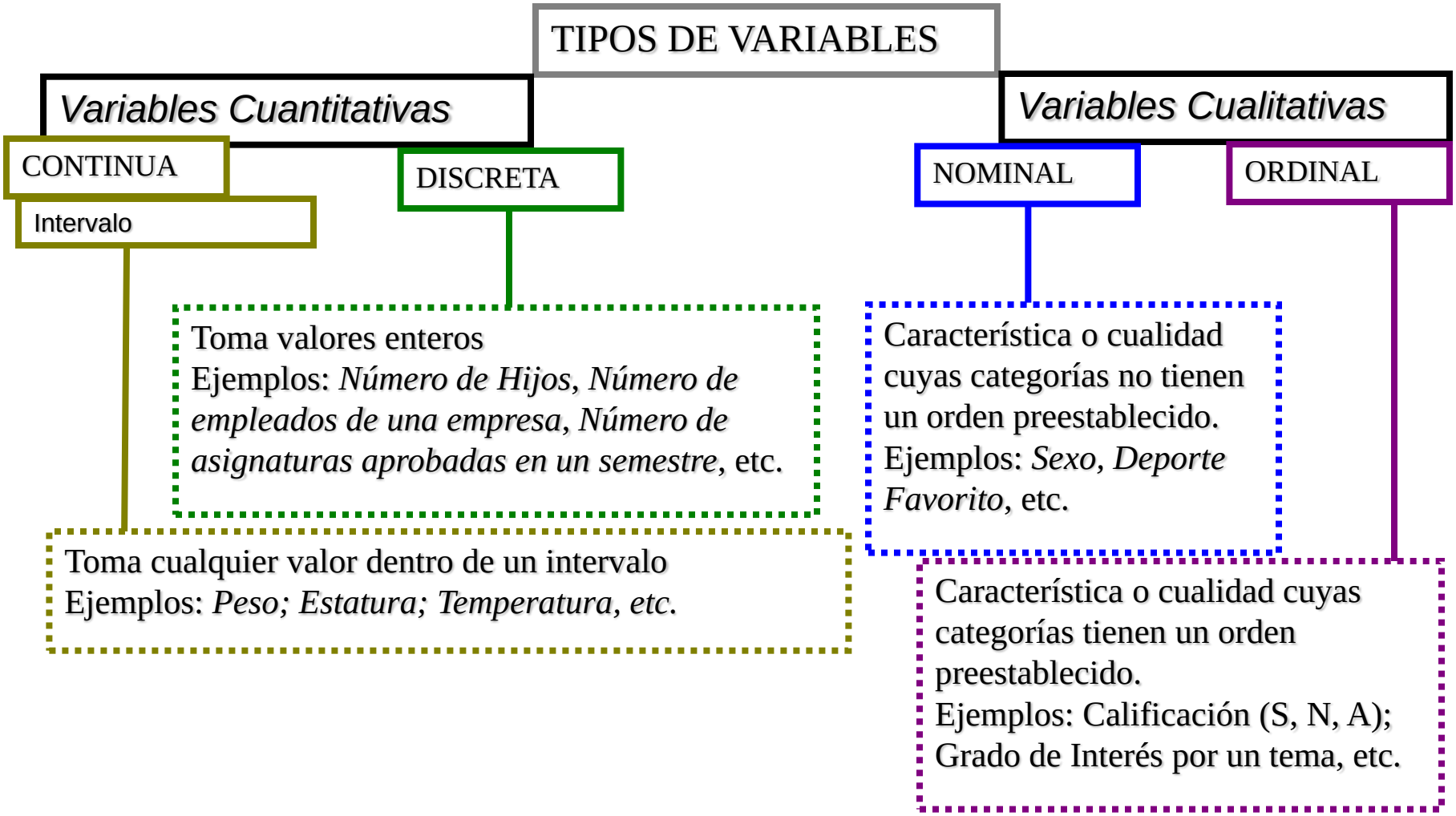
• VARIABLE: es lo que se va a medir y representa una característica de la UNIDAD DE ANÁLISIS.
• ¿QUIÉNES VAN A SER MEDIDOS?: Los sujetos u objetos o Unidades de Análisis de una Población o una Muestra

• POBLACIÓN : Es el total de unidades de análisis que son tema de estudio.

• MUESTRA: Es un conjunto de unidades de análisis provenientes de una población.

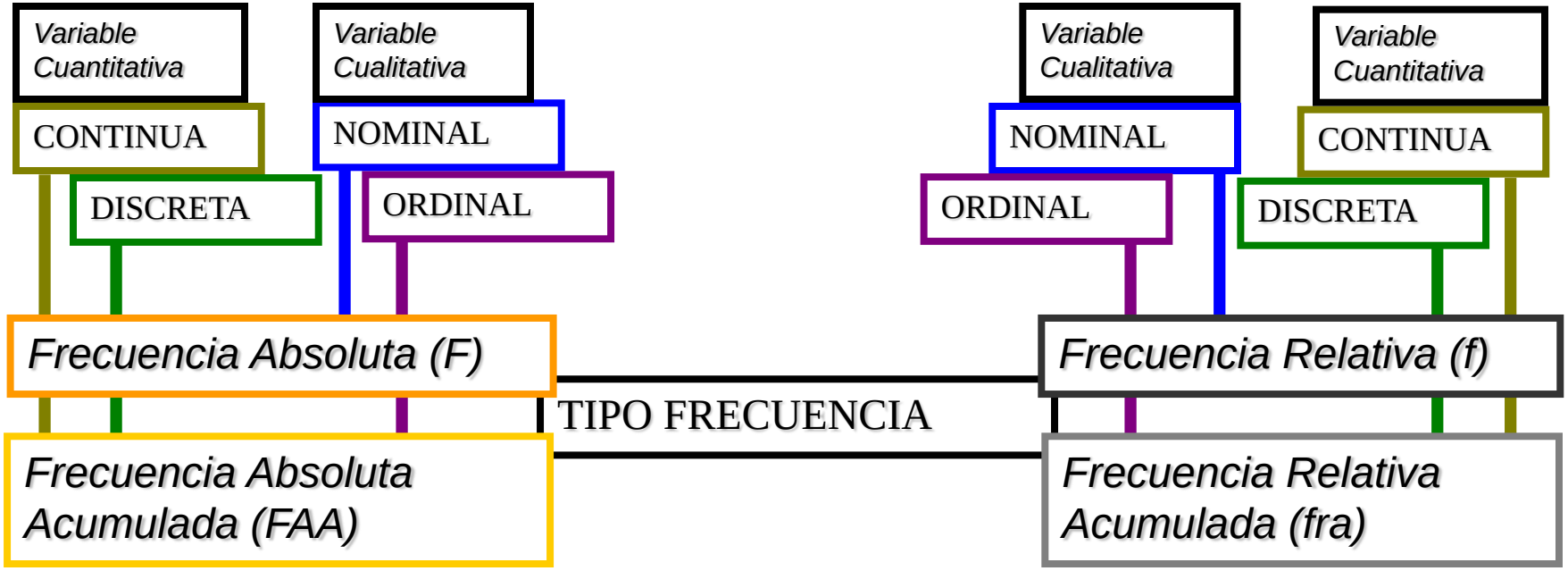


Variable: corresponde a la característica de la Unidad de Análisis



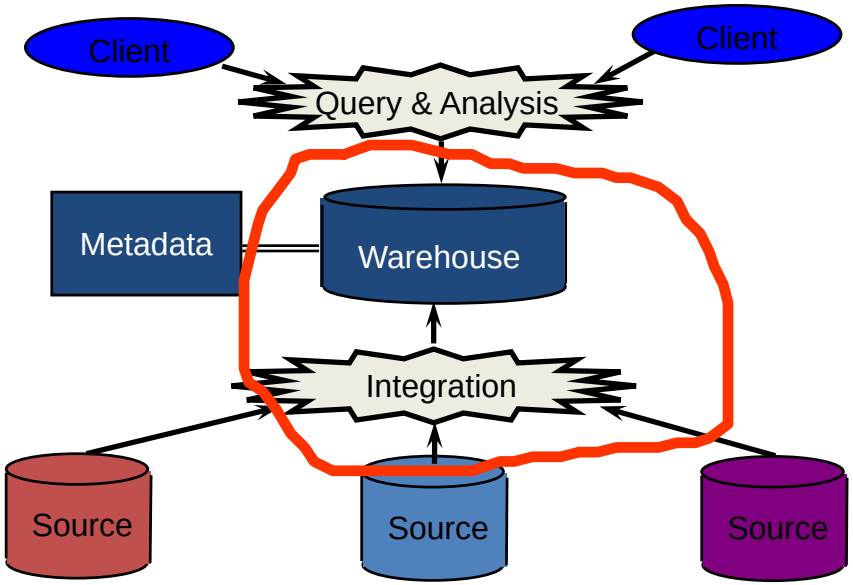
Unidad de Medida: Gramos o Kilos para la variable Peso; Grados C o F para Temperatura

Frecuencia: desde un conjunto de unidades, corresponde al Número o Porcentaje de veces que se presenta una característica.

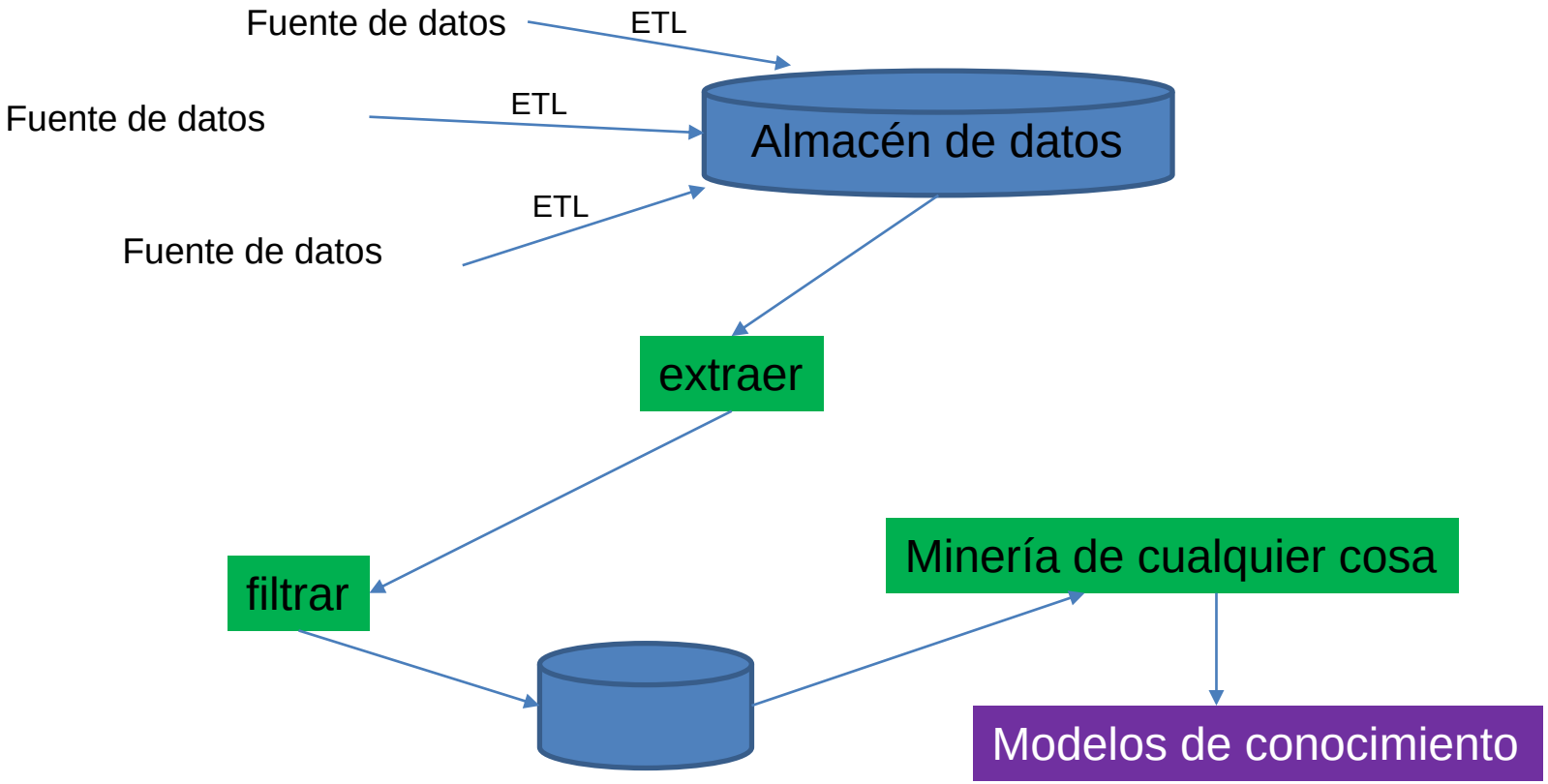


Integración

- Selección de los datos
- Transformación de datos
- Carga de datos



Descripción del flujo de trabajo de obtener desde fuentes de datos heterogéneas para procesar datos



Problemas: Calidad de los datos

- valores faltantes

cómo interpretar? ¿no disponible? usar el medio?

- valores duplicados: incluyendo partidos parciales (Jon Smith = John Smith?)
- inconsecuencia:

múltiples direcciones para la persona

- datos desactualizados
- uso inconsistente:

Significa "destino" de la primera etapa o de vuelo?

- valores atípicos:

sueldos que son negativos,

Problemas: Interoperabilidad

- ¿Cómo pueden compararse o combinarse datos de una base de datos con otra?
- ¿Qué pasa si los campos no son lo mismo, o no están presentes, o se usan de manera diferente?
- traducción / mapeo de términos estándares
  unidades como metros, o galones, etc.
  identificadores como SSN, UIN, ISBN
- Bases de datos "federadas": consultas que combinan información en varios servidores

"Limpieza de datos"

- completar los datos faltantes
- detectar y eliminar valores atípicos
  suavizado
  - eliminar ruido promediando valores juntos
- filtrado, muestreo
  - manteniendo solo valores representativos seleccionados
- extracción de características
  - p.ej. en una base de datos de fotos, ¿qué personas usan anteojos? Cuales tienen mas de una persona?

Preparación de los Datos

- **Recolección de datos**
  - Captura de la Información
- **Análisis**
  - Entender el contexto de la información
- **Preparación**
  - Hacer ciencia en los datos

http://www.youtube.com/watch?v=-xR5erOhkXo

Conocer los datos

Expertos de dominio deben ser consultados para explicar las anomalías, los valores perdidos, el significado de los números enteros que representan categorías en lugar de cantidades numéricas, y así sucesivamente.

La ciencia de datos es un procedimiento que **consume tiempo y requieren mucho trabajo**, pero que es absolutamente necesario para la AD con éxito.

Preparando la Entrada

Preparación de entrada para una investigación de AD suele **consumir la mayor parte del esfuerzo invertido en el proceso**.

Los datos deben pasar por **procesos de ensamblaje, integración, limpieza, agregación y preparación general**.

Recopilación de los datos en conjunto

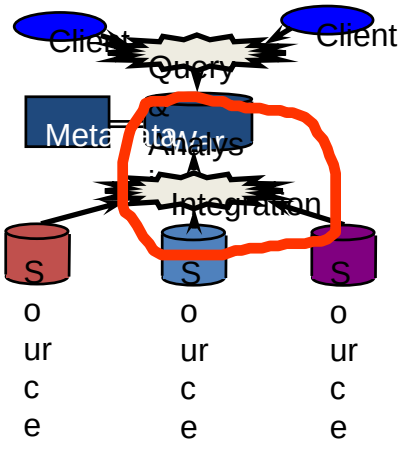
Proceso ETL

ETL (Extracción, Transformación y Carga)

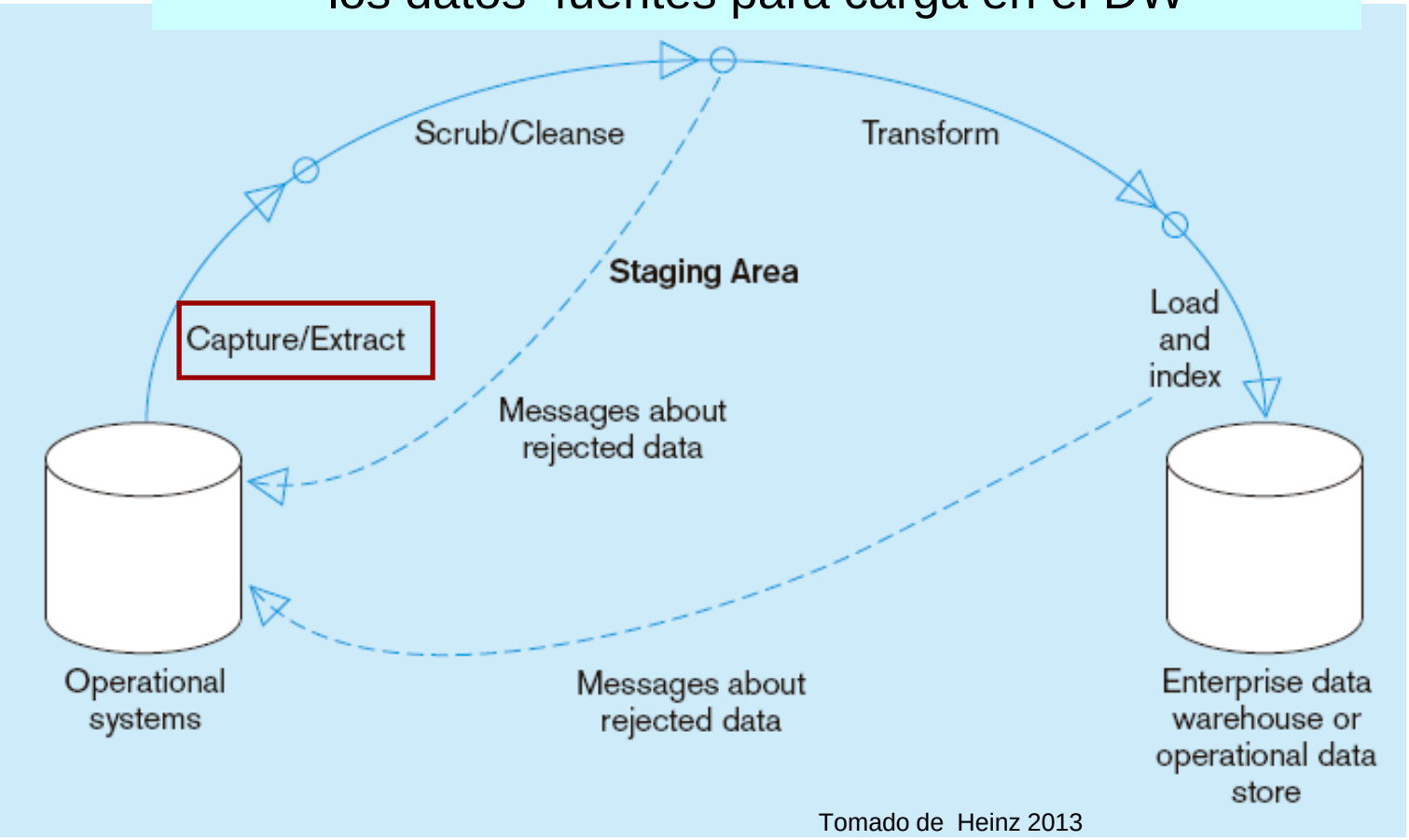
Extracción: Obtención de información de las distintas fuentes, tanto internas como externas.

Transformación: Filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: Organización y actualización de los datos y los metadatos en el DW.



Captura / Extrae... obtiene un subconjunto de los datos fuentes para carga en el DW



Tomado de Heinz 2013

Extracción estática: captura los datos en un momento puntual

Extracción Incremental: captura cambios que se van produciendo

Extracción

- Obtener datos de múltiples, heterogéneas fuentes externas
- Periódica
- Claves:
  - Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
  - No depender de la disponibilidad de los OLTP.
  - Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
  - Facilitar la integración de las diversas fuentes, internas y externas.

ETL: Técnicas de Monitoreo

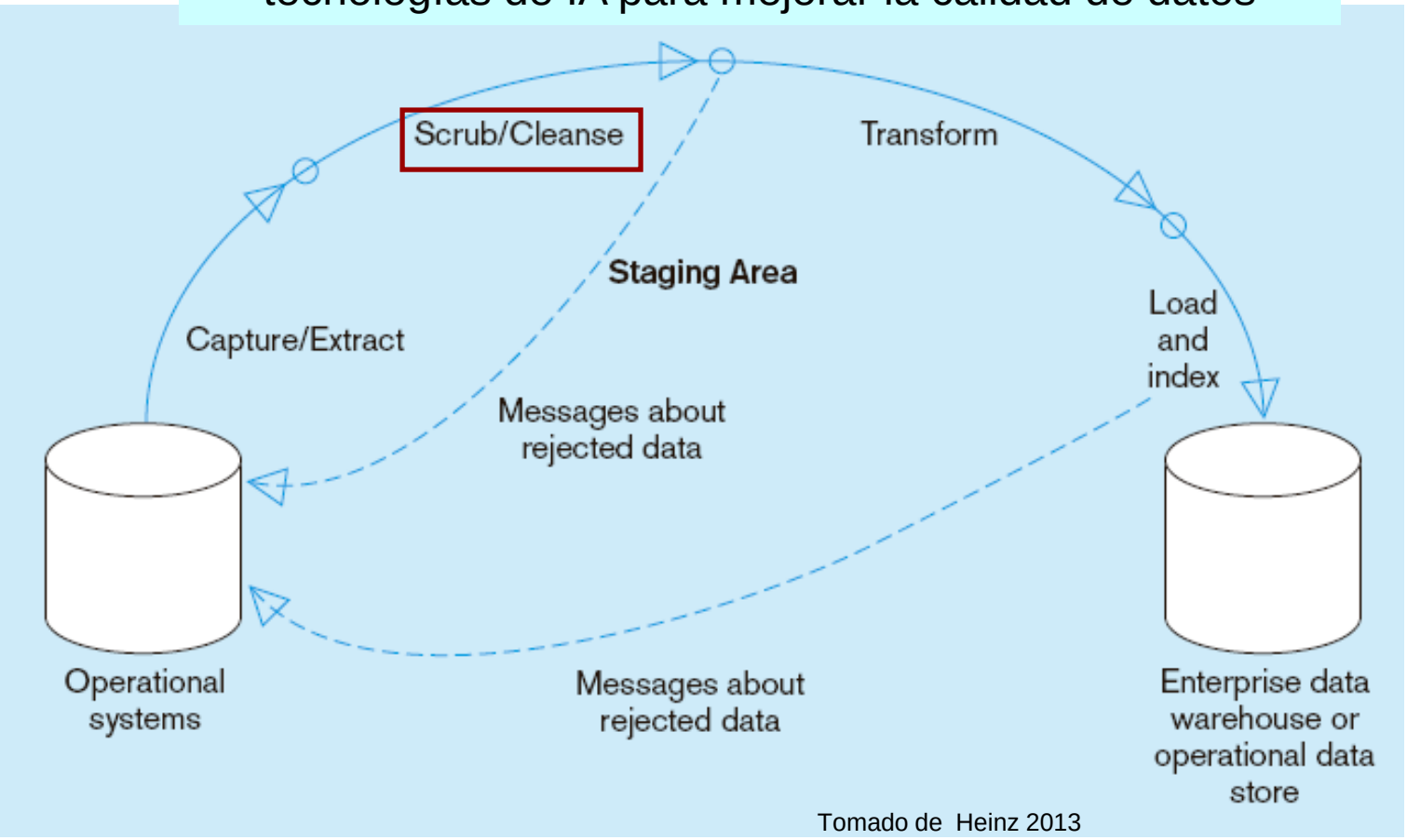
- Instantáneas cada cierto tiempo
- Disparadores de base de datos (reglas (triggers))
- Logs de registros
- Envío de datos
- Envío de transacciones
- Consultas a las BDs

Métodos de Extracción

- **Extracciones a granel**
  - Todo el DW se actualiza periódicamente
  - Pesado para las conexiones de red entre el origen y destino
  - Más fácil de configurar y mantener

- **Extracciones basadas en intercambio**
  - Sólo los datos que han sido recién insertadas o actualizadas en los sistemas de origen se cargan en el DW
  - Menos carga en la red pero requiere una programación más compleja para determinar cuando un nuevo registro de DW debe insertarse o cuando un registro DW tiene que actualizarse

Limpieza... utiliza reconocimiento de patrones y tecnologías de IA para mejorar la calidad de datos



Tomado de Heinz 2013

Solución de errores: faltas de ortografía, fechas erróneas, uso de campo incorrecto direcciones no coinciden, datos faltantes, datos duplicados, inconsistencias

También decodifica, reformatea, convierte, genera claves, fusiona, detecta errores registro, localiza datos faltantes

Limpieza

se refiere a una serie de procesos en los cuales la **calidad de los datos es mejorada**, enfrentando los problemas mencionados como datos mal capturados, anómalos y vacíos, ya sea por características obvias que el dato no cumple con ciertos parámetros del estándar, o porque el experto del proceso ya tiene identificado anomalías comunes en el almacenamiento de los datos.

- normalización de formatos,
- remoción de anomalías,
- corrección de errores
- eliminación de duplicados.

Limpieza de datos

Sistemas de origen tiene "datos sucios" que deben ser limpiados

Razones de datos “sucios"

- valores ficticios
- La falta de datos
- datos encriptados
- datos contradiciéndose
- El uso inapropiado de datos
- Violación de Reglas de Negocio
- Claves principales reutilizados
- Identificadores no únicos

detectar errores en los datos y rectificarlos

Limpieza

Ejemplos anomalías que presenta la base de datos:

- Unidades de las entradas
- Abreviaciones
- Convenciones de nombres
- Representaciones diferentes
- Variaciones de Ortografía
- Elementos repetidos
- Datos no guardados

Para buscar anomalías :

- Estudiar la representación de cada una de las variables.
- Buscar anomalías de representación.
- Definir alguna estrategia de limpieza para erradicar dichas
- Realizar las operaciones con un software para limpieza de datos.

Pasos en la Limpieza de datos

1. Análisis (parsing)
2. Corrección
3. Estandarización
4. Mapeo (matching)
5. Consolidación

Análisis (Parsing)

Localiza e identifica **elementos individuales** en los archivos de origen y luego los aísla.

- **Ejemplos**
  - Análisis del primer nombre, segundo nombre y apellido;
  - Analizar número y nombre de la calle;
  - Analizar la ciudad y el estado.

Corregir

Corrige los componentes de **datos individuales** utilizando algoritmos sofisticados y fuentes de datos secundarias.

Acciones típicas con Datos Anómalos (Outliers):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.

Acciones contra Datos Faltantes (Missing Values):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Esperar hasta que los datos

Estandarización

- Aplica **rutinas de conversión** para transformar los datos en formatos preferidos (y coherentes), utilizando reglas de estándarización.

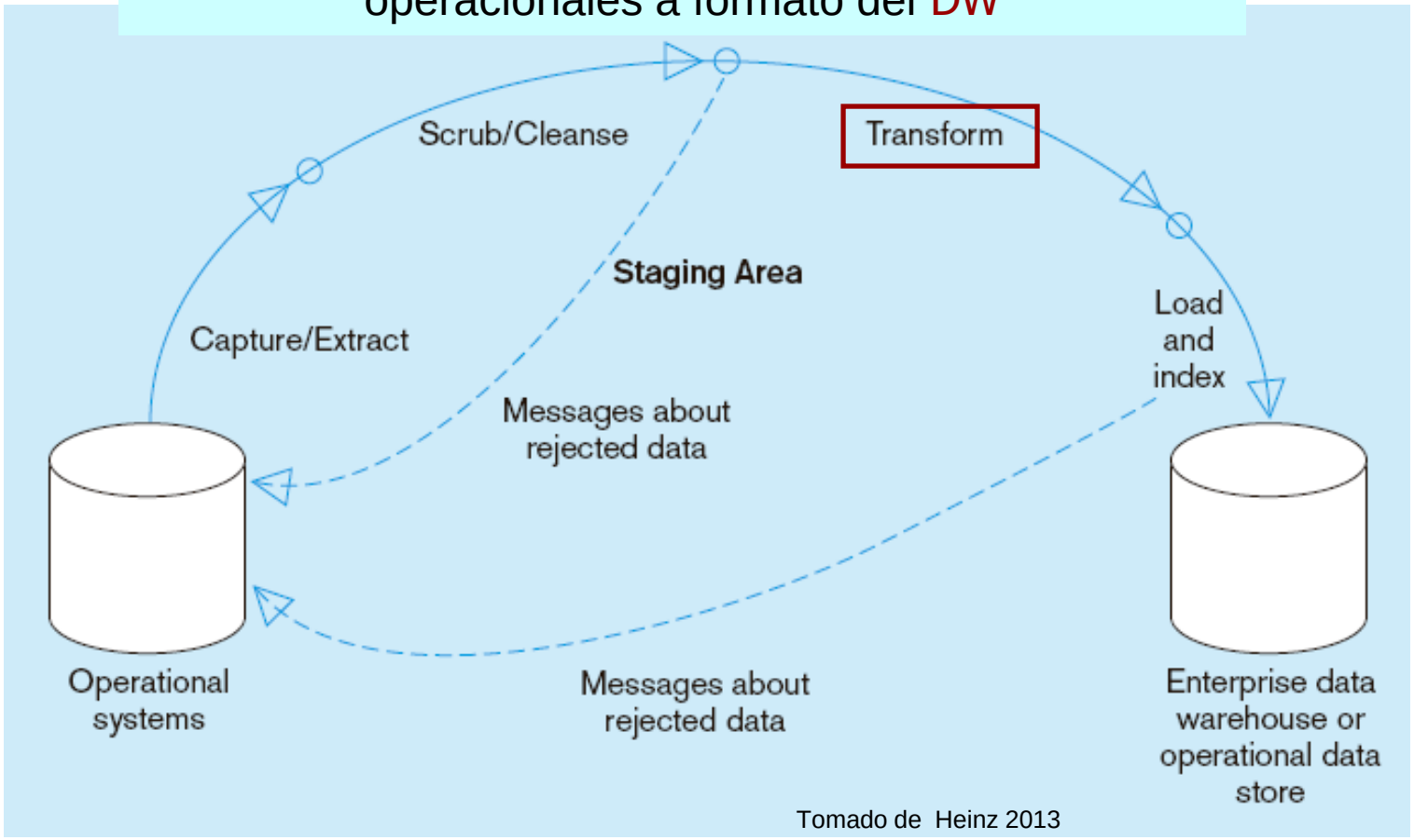
- **Ejemplos:** la sustitución de un apodo, uso de un nombre de calle preferido, etc.

Mapeo (matching)

- búsqueda de **coincidentes en registros** en los datos analizados, corregidos y estandarizados, basados en reglas predefinidas para eliminar la duplicación.

- **Ejemplos:** identificación de nombres y direcciones similares.

Transforma… convierten los datos desde las BD operacionales a formato del DW



Tomado de Heinz 2013

A nivel de registro:

- Partición de Datos (selección)
- Juntar datos (combinación)
- Resumir datos (agregación)

A nivel de campo:

- de un solo campo: de un campo a un campo
- multi-campo: de muchos campos a uno, o uno a muchos campos

Transformación de datos

- Transforma los datos de acuerdo con reglas y normas establecidas

- Clásicos problemas:
  - Codificación.
  - Medida de atributos.
  - Convenciones de nombramiento.
  - Fuentes múltiples,

ordena **calcula**

resume **consolida**

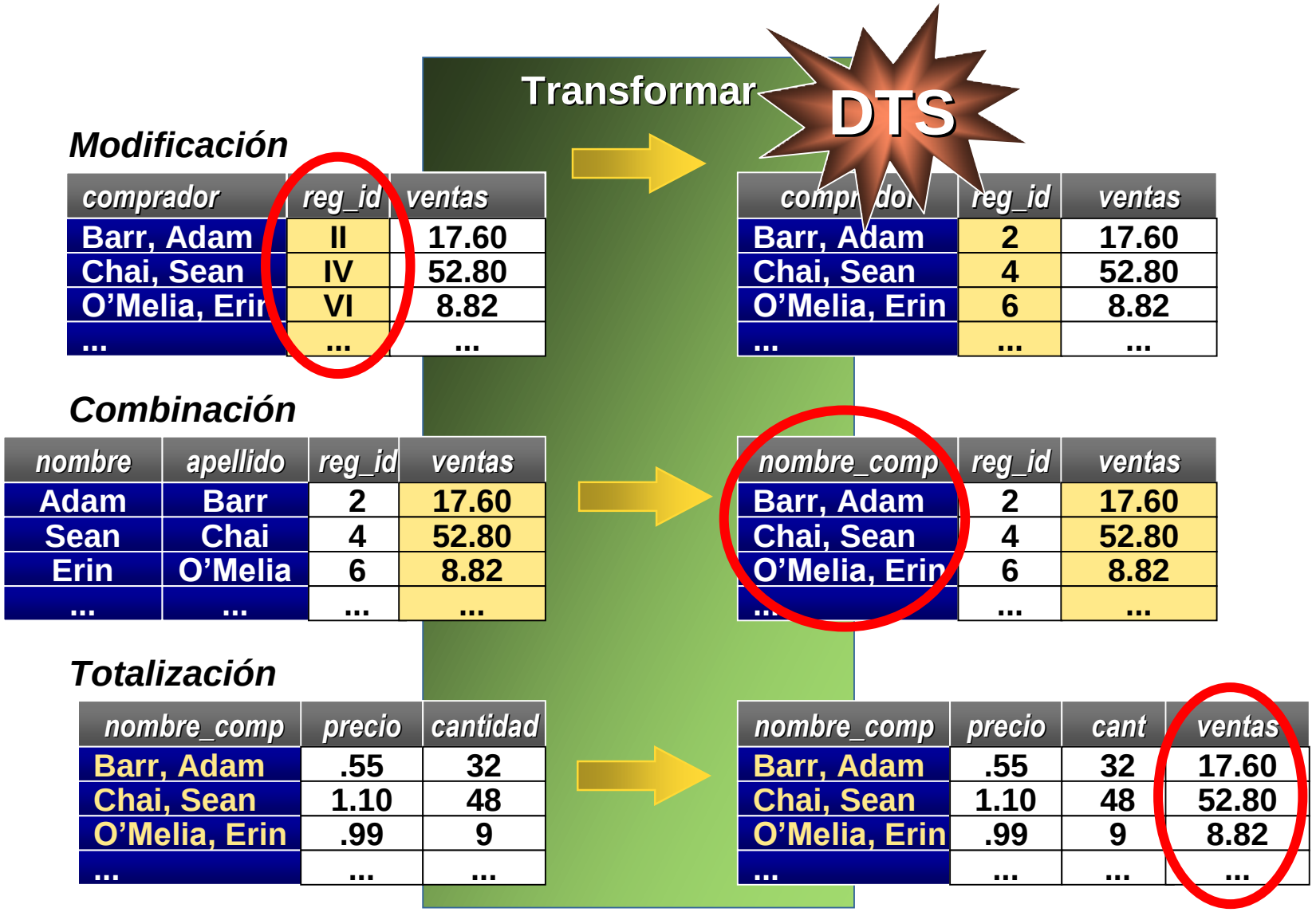
Transformación

En esta etapa se **transforman las variables de entrada en nuevas variables de interés**, esto se realiza a través de diversos métodos, los cuales se deben escoger en caso de ser pertinente alguna transformación de alguna de las variables.

Una transformación de variables puede ser la combinación entre variables

- concatenación de cadenas,
- multiplicación entre variables,
- otras operaciones aritméticas, etc.

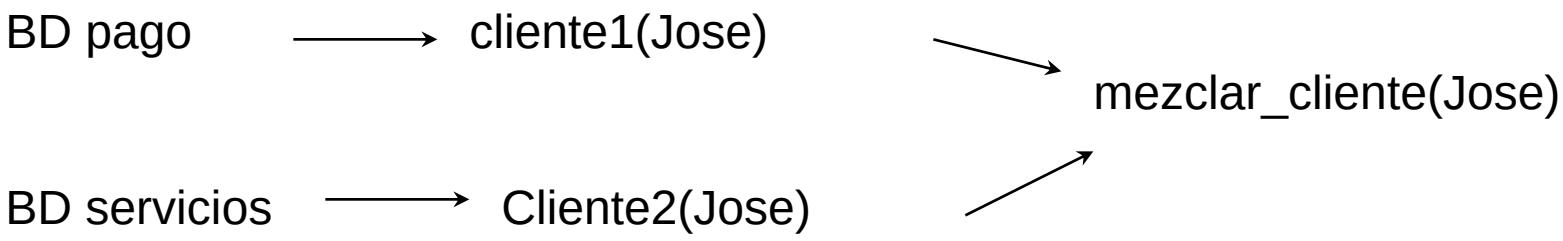
Proceso ETL; Consideraciones



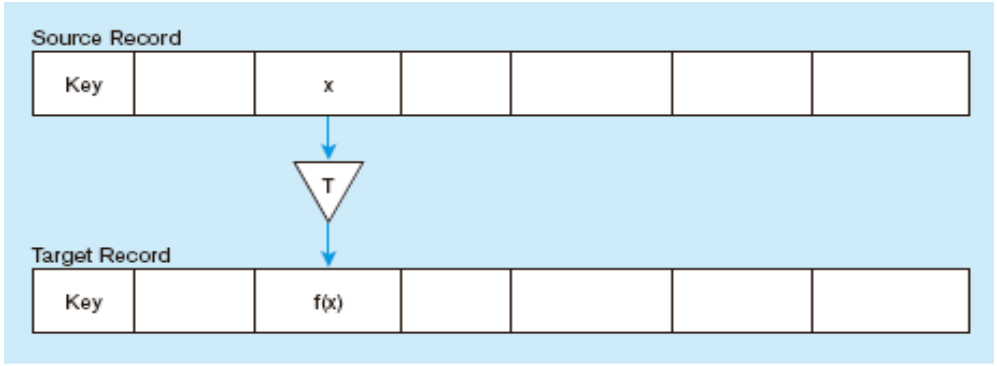
ETL: Transformación de datos

Posibles tareas

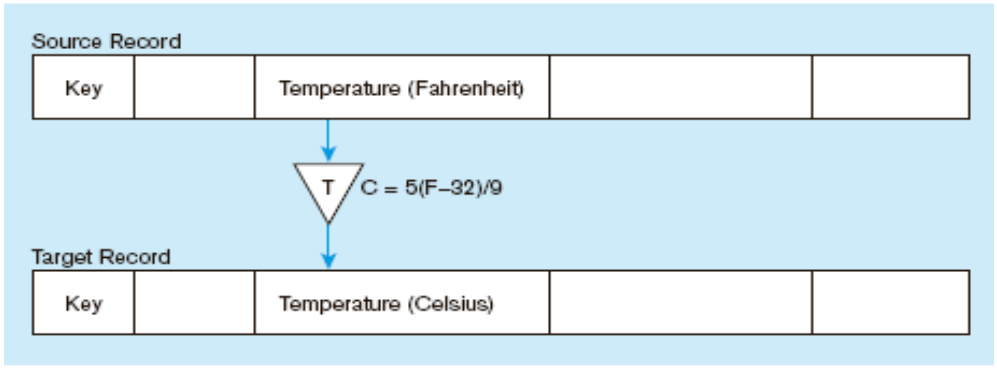
- **Migrar** (por ejemplo, yen a dólares)
- **Refinar**: utilizar el conocimiento específico de dominio (por ejemplo, números de seguro social)
- **Fusionar** (por ejemplo, lista de correo con la de clientes)



Transformación de un solo campo



Una función de transformación traduce los datos de una forma antigua a una nueva



Transformación algorítmica: utiliza una fórmula o expresión lógica

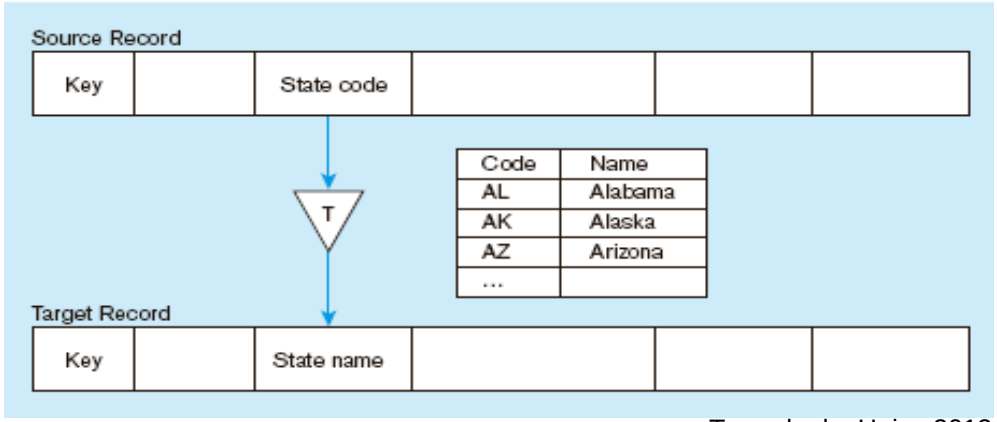
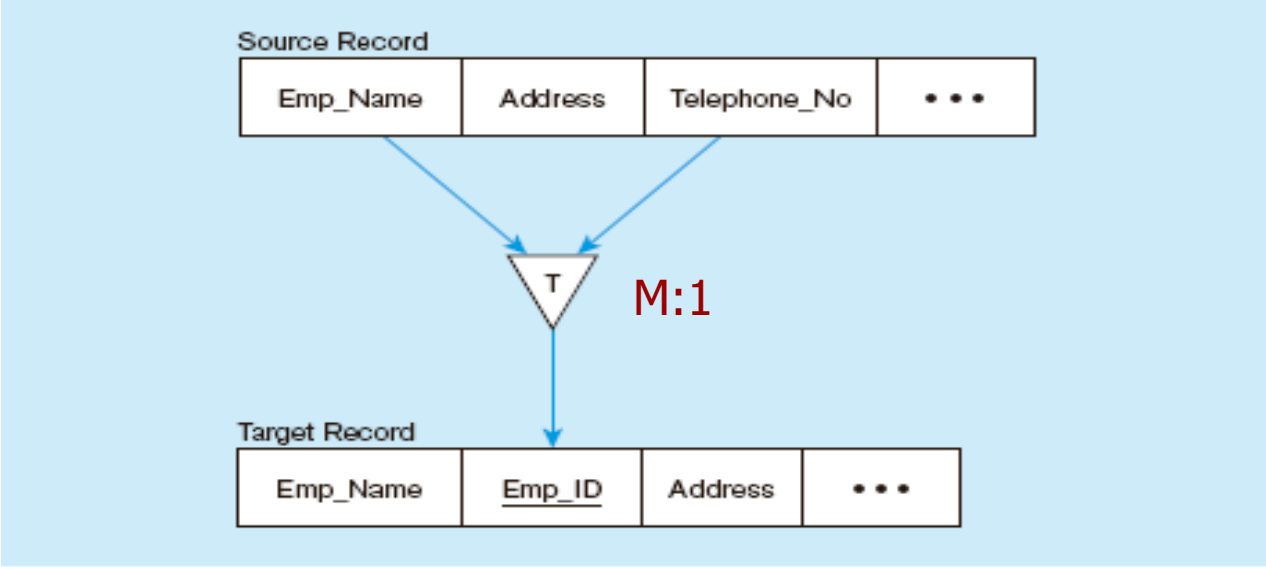
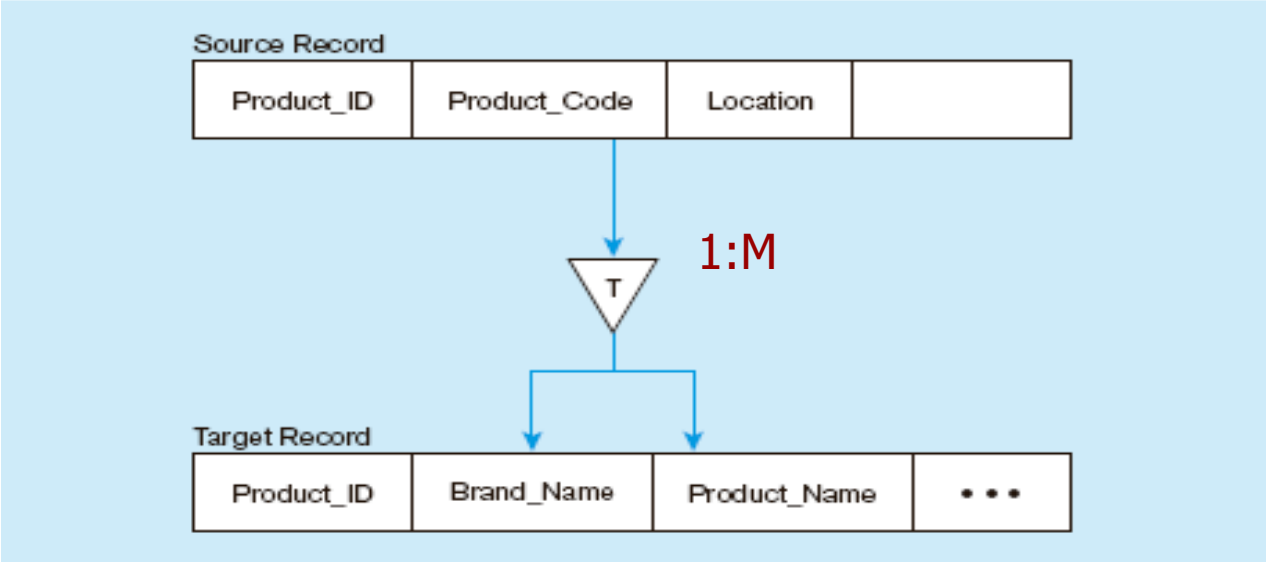


Table lookup: utiliza una tabla separada basada en códigos

Tomado de Heinz 2013

Transformación multicampos





Proceso Transformación

- Estudiar las representaciones de cada una de las variables
- Identificar las representaciones que se puedan transformar en otra representación más conveniente o fácil de utilizar para AD.
- Ordenar dichas transformaciones que se desean aplicar en una tabla, para observar las equivalencias
- Aplicar la transformación con el software seleccionado
- Identificar las variables que potencialmente se pueden normalizar
- Definir la función(es) de normalización para cada una de las variables seleccionadas en el paso anterior y ordenarla en tablas.
- Aplicar la función(es) de normalización en las variables seleccionada
- Combinar variables por un método seleccionado tal como el PCA (del inglés *Principal Component Analysis*) que es considerado también un método para reducción de variables.
- Describir en tablas cada una de las transformaciones realizadas.

Reducción

Consiste en **decidir qué datos deben ser utilizados para el análisis**. El criterio que se sigue para realizar reducción de variables incluye la relevancia con respecto a los objetivos que se persiguen, y limitaciones técnicas tales como los volúmenes máximos de datos o bien tipos de datos concretos.

Así que en este paso se reduce la cantidad de variables a sólo las necesarias para modelar el proceso en estudio.

- Realizar análisis estadísticos para reducir variables que posean una alta relación lineal, como por ejemplo un análisis de correlación.
- Identificar las posibles variables que se pueden reducir.
- Justificar la reducción de las mismas
- Construir la nueva vista minable con las nuevas variables reducidas

Datos Dispersos

- La mayoría de los atributos tienen un valor de 0
  - Los registros de datos de la cesta de compras realizadas por los clientes de los supermercados

- Puede ser poco práctico representar cada elemento de una matriz dispersa de forma explícita:

  - 0, 26, 0, 0, 0, 0, 63, 0, 0, 0, "clase A"
  - 0, 0, 0, 42, 0, 0, 0, 0, 0, 0, "clase B"

En cambio, los atributos no nulos pueden ser explícitamente identificado por el número de atributo y su valor declarado:

{1 26, 6 63, 10 "clase A"}
{3 42, 10 "clase B"}

Los tipos de atributo

Existen dos tipos básicos de atributos nominales y numéricos.

Los atributos string y date son atributos nominales y numéricos, respectivamente.

Aunque previamente las cadenas pueden ser convertidas en la forma numérica como un vector de palabras

Valores Perdidos

Los Valores perdidos son frecuentemente indicados por fuera del rango de entradas, tal vez un número negativo (-1).

A veces, diferentes tipos de valores perdidos se distinguen (por ejemplo, valores desconocidos: sin grabar vs irrelevante) y que pueden estar representados por diferentes números enteros negativos (-1, -2)

Pueden ocurrir por varias razones:

- Equipos de medición funcionando de manera incorrecta
- Los cambios en el diseño experimental durante la recolección de datos
- Los entrevistados en una encuesta pueden negarse a responder a ciertas preguntas
- En un estudio arqueológico, un espécimen tal como un cráneo pueden ser dañados de forma que algunas variables no se pueden medir.

Valores inexactos

La base de datos debe ser revisada cuidadosamente.

Cuando es recolectada la información, es probable que no importe si hay campos en blanco o sin restricción en los archivos.

Si la misma base de datos es utilizada para AD, los errores y omisiones de inmediato comienzan a tomar significancia.

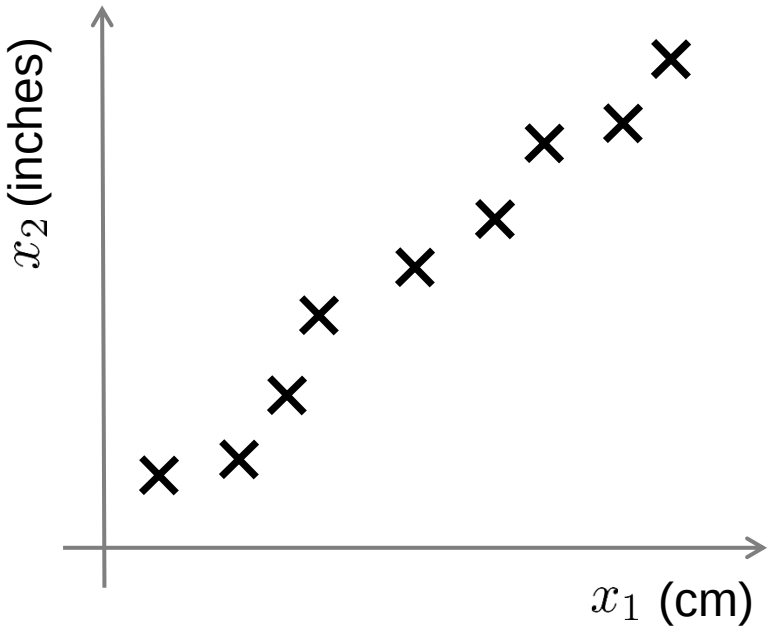
Por ejemplo:

- Datos duplicados son una fuente de error
- Datos que se convierten en obsoletos, hay que considerar si los datos a usar en minería son todavía validos o actuales

Compresión de los datos

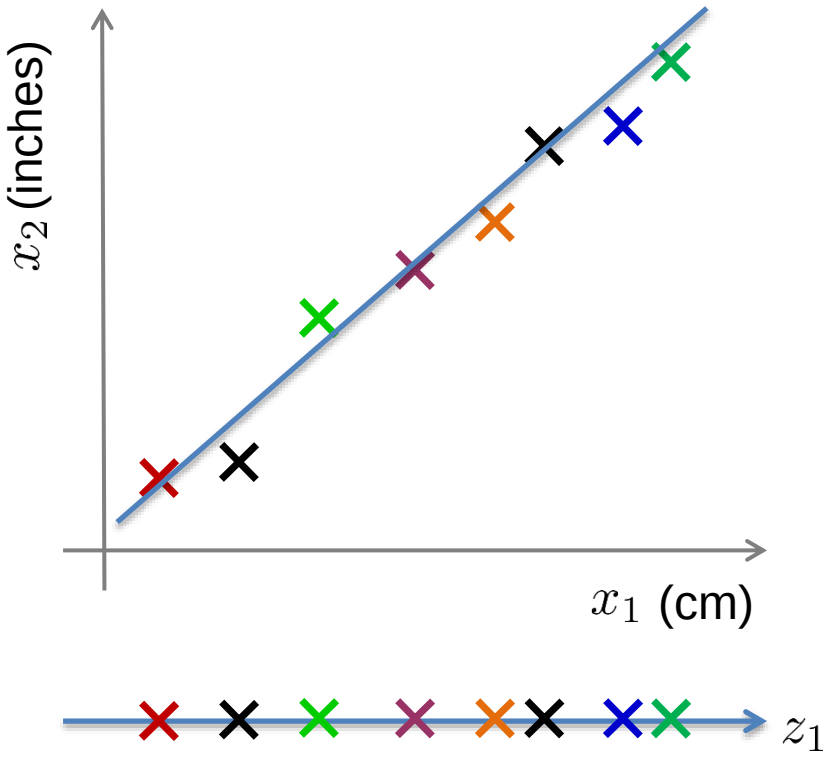
compresión de datos es la reducción del volumen de datos tratables para representar una determinada información empleando una menor cantidad de espacio.

Data Compression



Reduce data from 2D to 1D

Data Compression



Reduce data from 2D to 1D

$$x^{(1)} \in R^2 \rightarrow z^{(1)} \in R$$

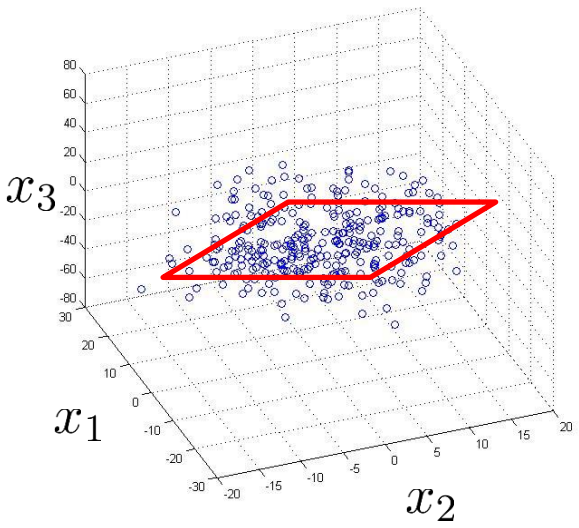
$$x^{(2)} \in R^2 \rightarrow z^{(2)}$$

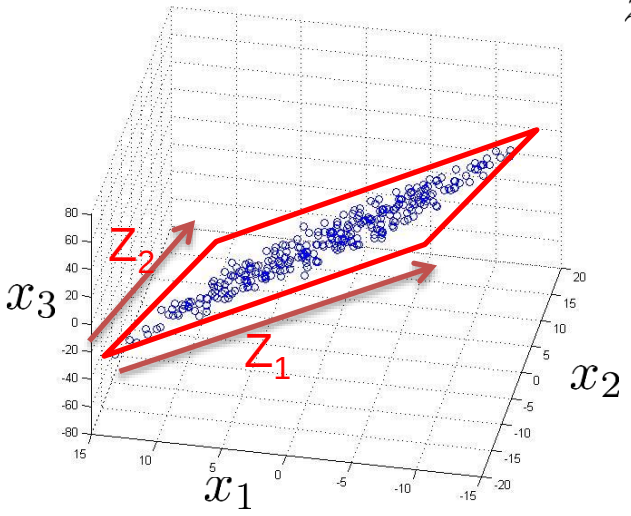
$$\vdots$$

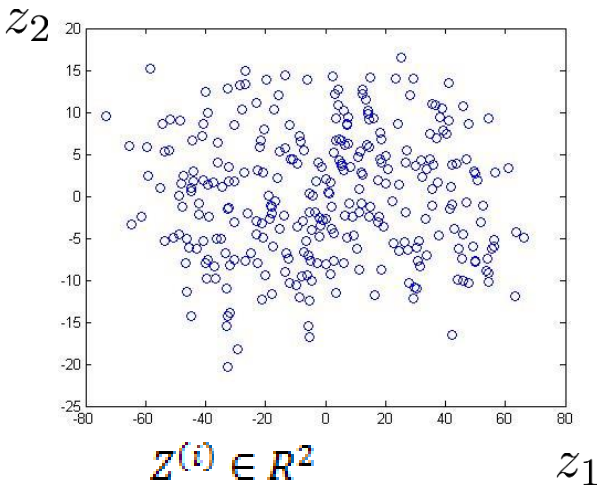
$$x^{(m)} \in R^2 \rightarrow z^{(m)} \in R$$

Data Compression

Reduce data from 3D to 2D







$X^{(i)} \in R^3$

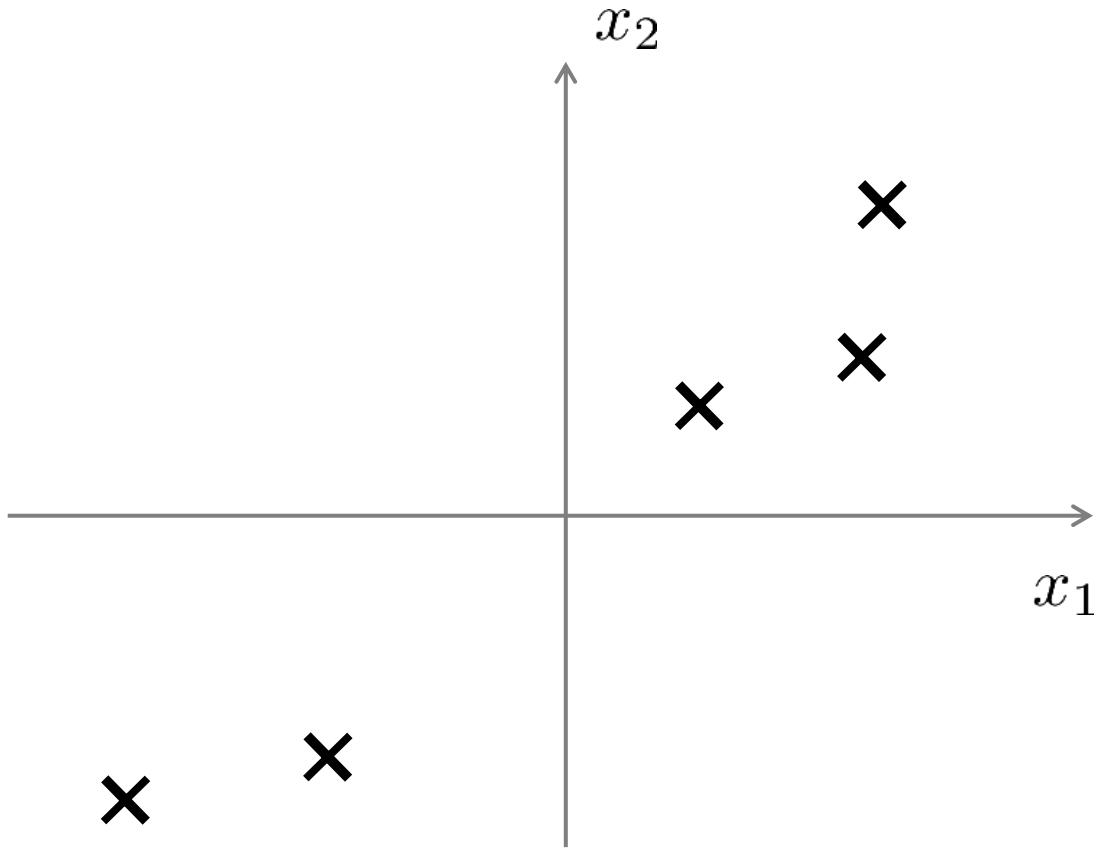
$Z^{(i)} \in R^2$

Formulación del Problema PCA (Principal Component Analysis)

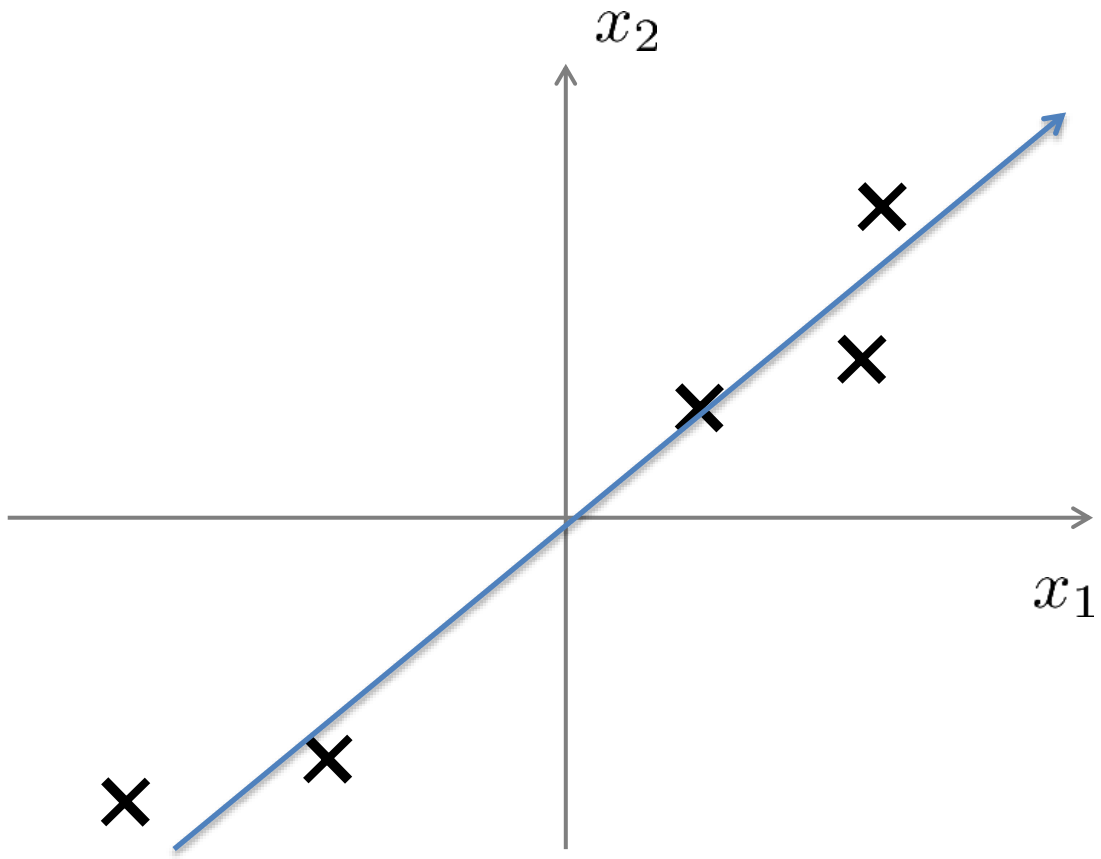
El **análisis de componentes principales** (en español **ACP**) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos.

Intuitivamente la técnica sirve para hallar las causas de la variabilidad de un conjunto de datos y ordenarlas por importancia.

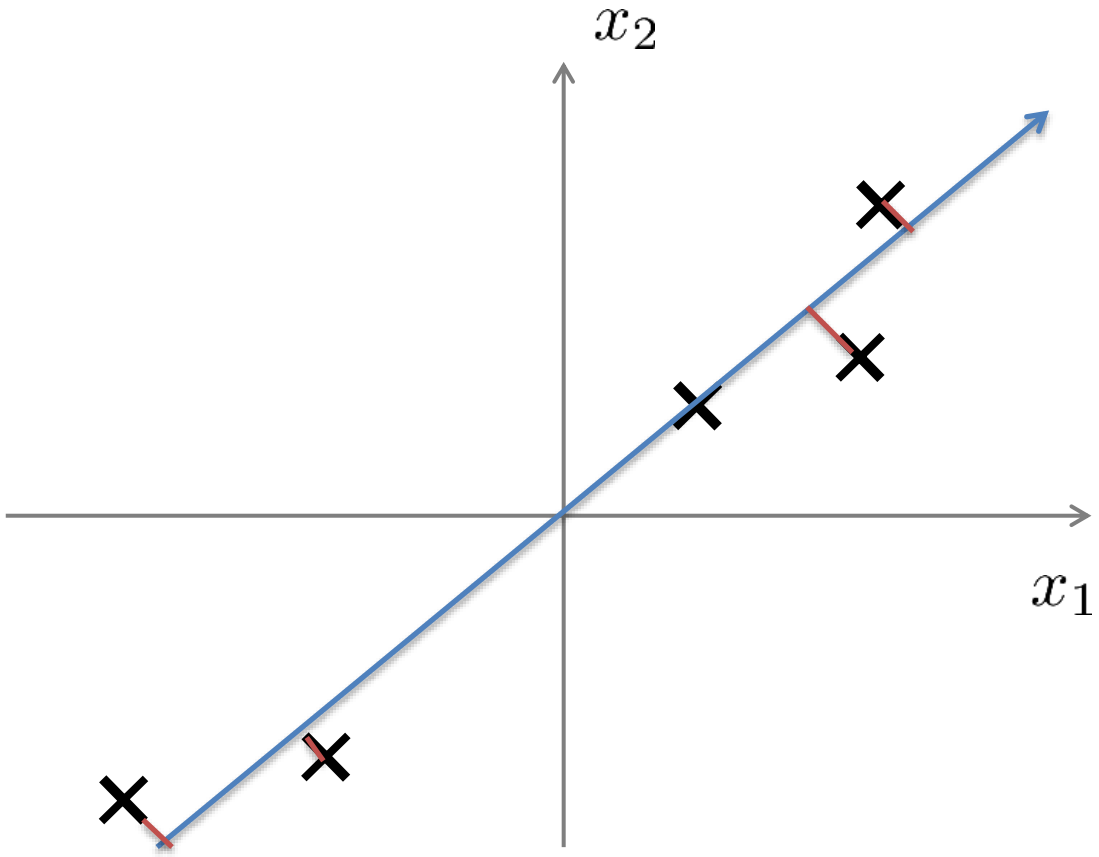
Formulación del Problema PCA (Principal Component Analysis)



Formulación del Problema PCA (Principal Component Analysis)



PCA (Principal Component Analysis) Error de proyección



PCA (Principal Component Analysis) Error de proyección

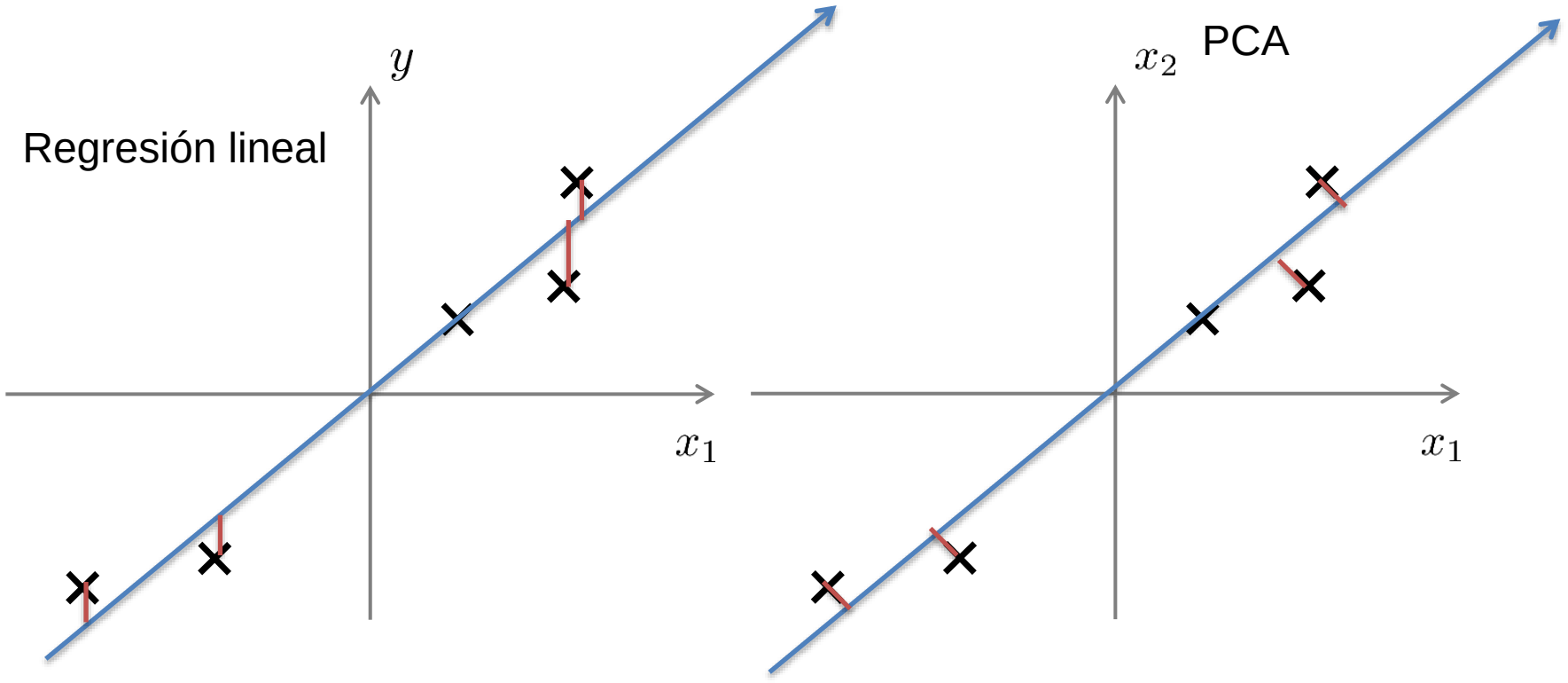
$$u^{(1)} \in \mathbb{R}^n$$

Para reducir de 2- dimensiones a 1-dimensión: Buscar un vector
Para proyectar la data, el cual minimice el error de proyección.

$$u^{(1)}, u^{(2)}, \ldots, u^{(k)}$$

Para reducir de n-dimensiones a K-dimensiones: Buscar K vectores
Para proyectar la data, los cuales minimicen el error de proyección.

PCA no es regresión lineal



Data preprocessing

Training set: $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$

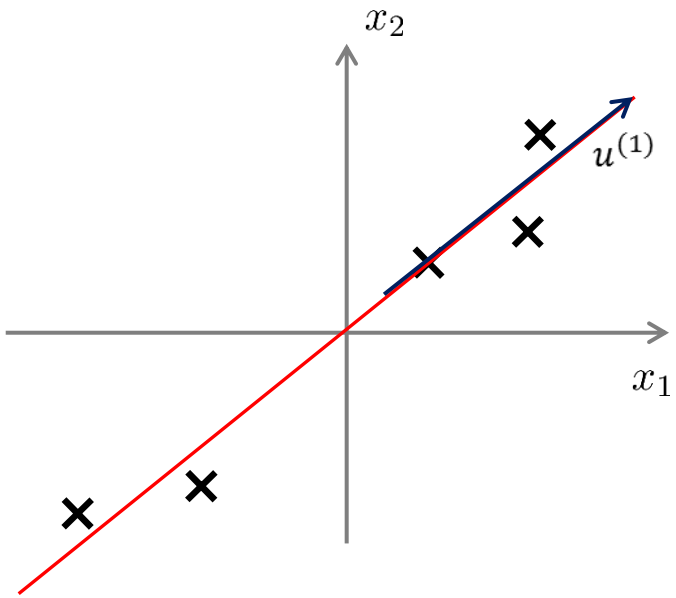
Preprocessing (feature scaling/mean normalization):

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

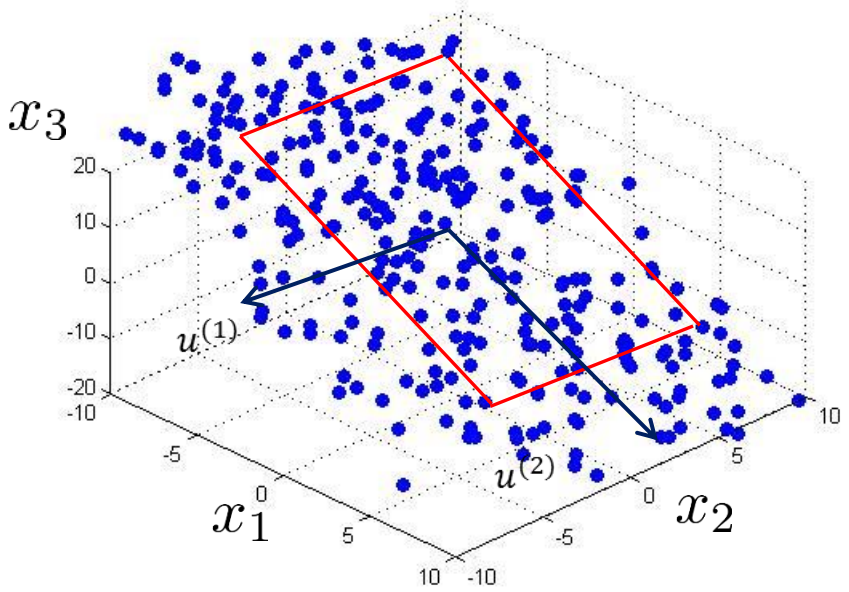
Replace each $x_j^{(i)}$ with $x_j - \mu_j$

If different features on different scales (e.g., $x_1 =$ size of house, $x_2 =$ number of bedrooms), scale features to have comparable range of values.

Principal Component Analysis (PCA)



Reduce data from 2D to 1D



Reduce data from 3D to 2D

Principal Component Analysis (PCA)

Reduce data from n-dimensions to k-dimensions

Compute "covariance matrix":

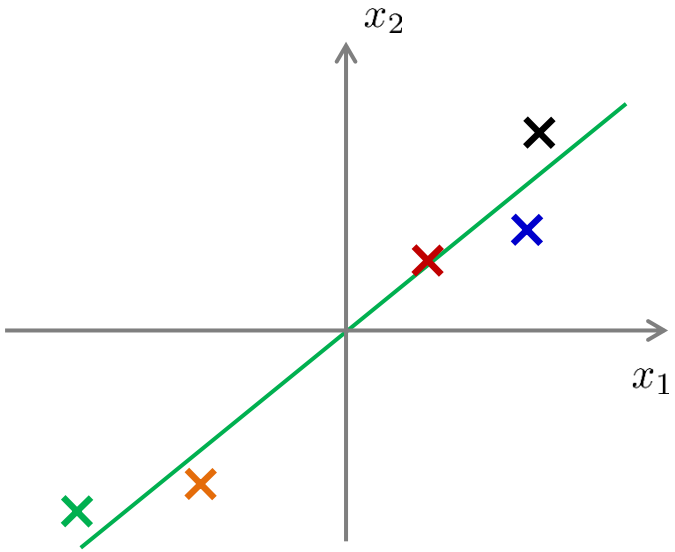
$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T \longrightarrow \text{Sigma}$$

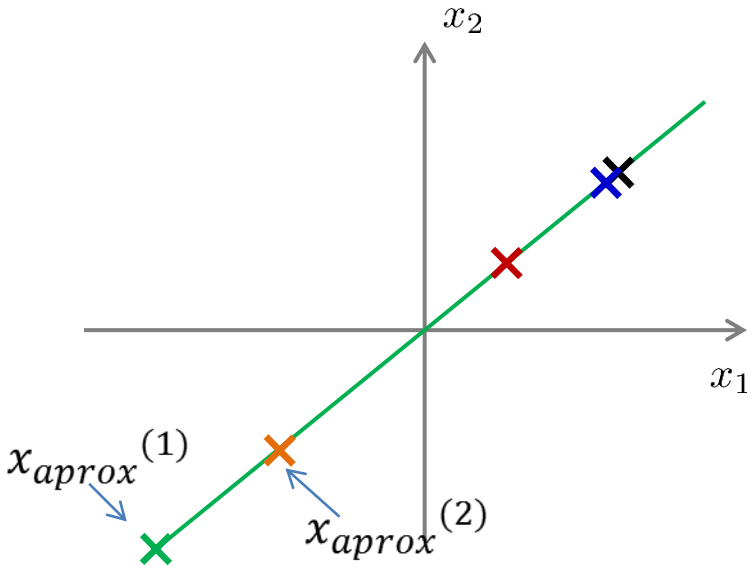
Compute "eigenvectors" of matrix Σ:

```
[U,S,V] = svd(Sigma);
```

El ACP comporta el cálculo de la descomposición en autovalores de la matriz de covarianza,

Reconstrucción





$$z = U_{reduce}^{T} x$$

$$X_{aprox} = U_{reduce} . Z$$

Formato ARFF

El formato ARRF, acrónimo de *Attribute-Relation File Format*. Este formato está compuesto por una estructura claramente diferenciada en tres partes:

1. **Cabecera.** Se define el nombre de la relación. Su formato es el siguiente:

    @relation <nombre-de-la-relación>

Donde <nombre-de-la-relación> es de tipo String*. Si dicho nombre contiene algún espacio será necesario expresarlo entrecomillado.

Formato ARFF

2. **Declaraciones de atributos.** En esta sección se declaran los atributos que compondrán nuestro archivo junto a su tipo. La sintaxis es la siguiente:

@attribute <nombre-del-atributo> <tipo>

3. **Sección de datos.** Declaramos los datos que componen la relación separando entre comas los atributos y con saltos de línea las relaciones.

@data

4,3.2

Formato ARFF

“arff” es la extensión del nombre del archivo

FILE FORMAT

@relation RELATION_NAME

@attribute ATTRIBUTE_NAME ATTRIBUTE_TYPR
@attribute ATTRIBUTE_NAME ATTRIBUTE_TYPR
@attribute ATTRIBUTE_NAME ATTRIBUTE_TYPR
@attribute ATTRIBUTE_NAME ATTRIBUTE_TYPR

@data
DATAROW1
DATAROW2
DATAROW3

Formato ARFF

archivo “arff”

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

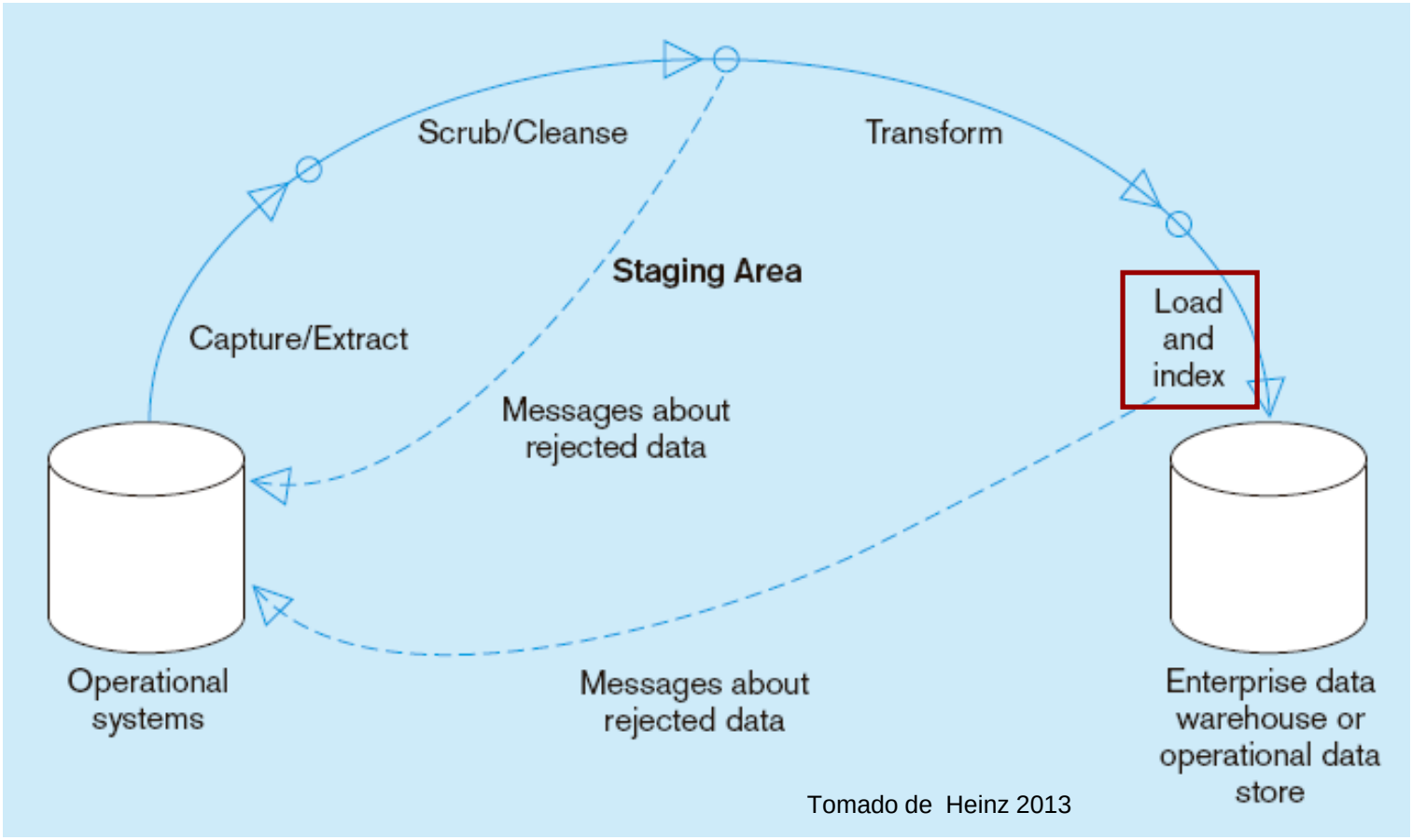
Formato ARFF

```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rainy, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 71, 91, true, no
```

Cargar/Indizar… Transforma datos y crea índices



Tomado de Heinz 2013

Modo de Actualización 1: reescritura masiva de datos en destino a intervalos periódicos

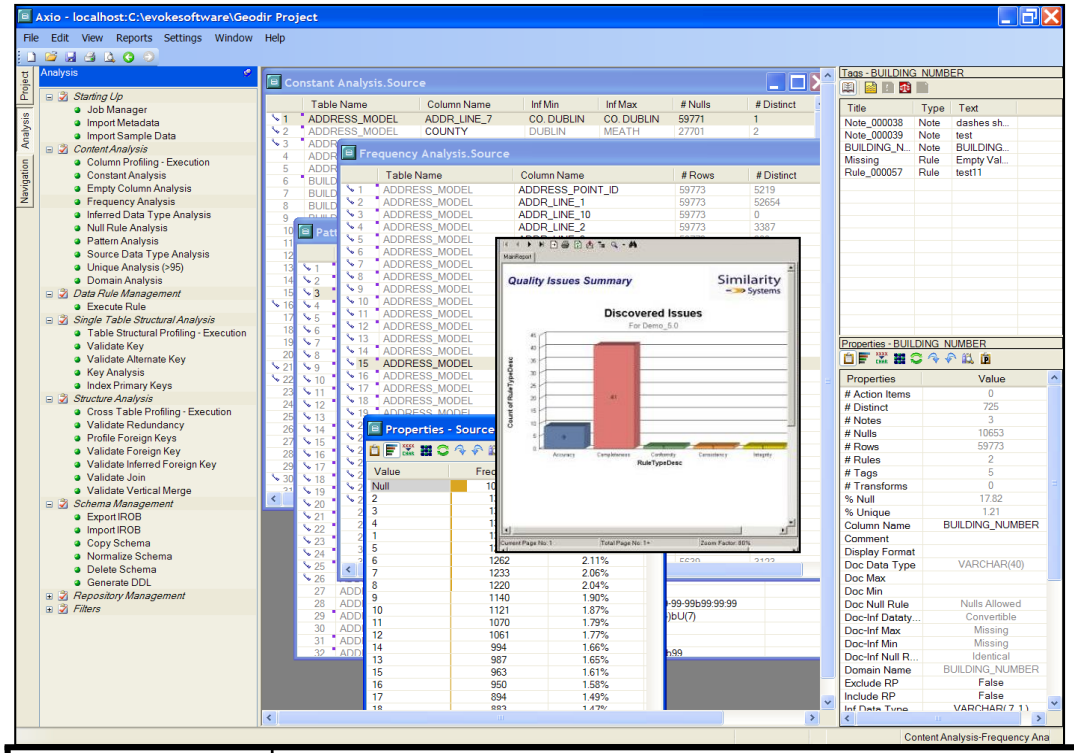
Modo de actualización 2: solamente cambios en los datos de origen se escriben en el warehouse

Carga de Datos

- Los datos físicamente se almacena en el almacén de datos
- La carga ocurre en una "ventana de carga“
- La tendencia cada vez mayor es actualizaciones en tiempo real

Preparación de los Datos

Perfil de los datos



Experto IN	Perfila y etiqueta anomalías
Ùsuario	Revisa anomalías

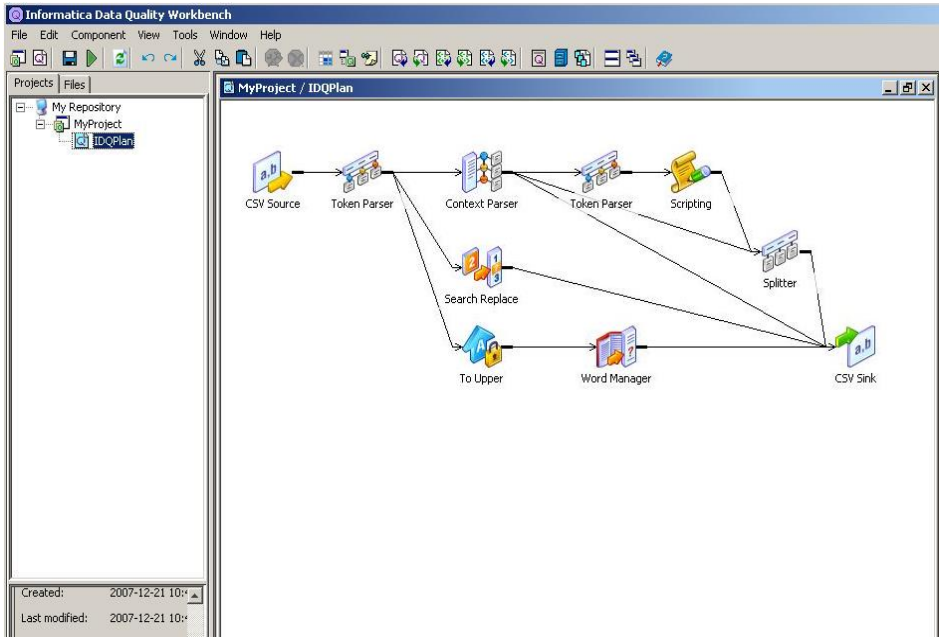
Catálogo de Problemas

- Completos
- Conformes
- Consistentes
- Precisos
- Duplicados
- Dependencias
- Correctas especificaciones y transformación
- **Íntegros**

Preparación de los Datos:

Diseñar e implementar reglas de calidad de datos

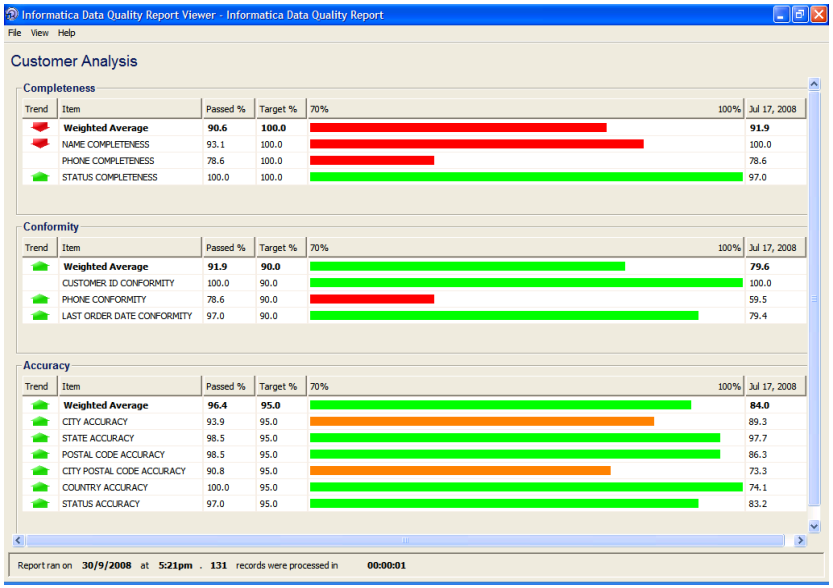
- Estandarizar los valores
- Datos faltantes o incorrectos
- Enriquecer los datos
- Identificar y consolidar los registros duplicados

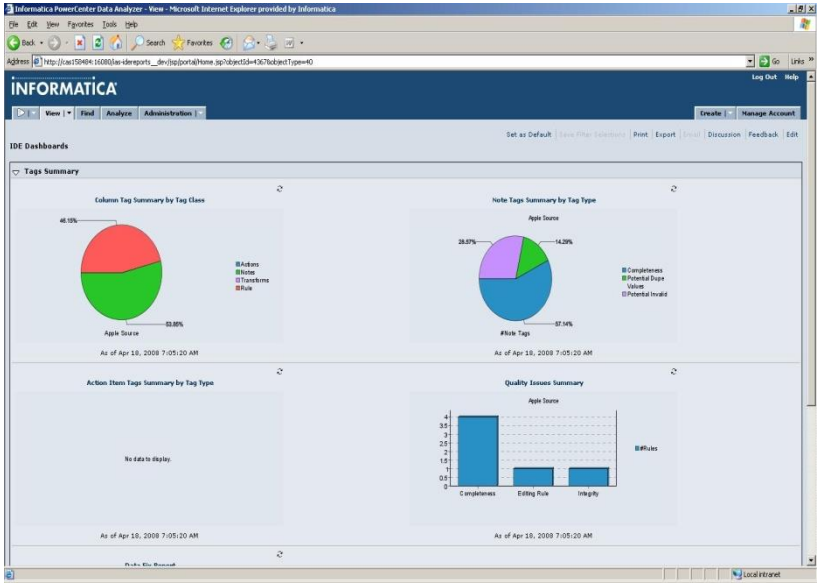


Experto IN	**Asiste con el diseño de reglas de cálidad**
Usuarios	**Crea reglas de cálidad de datos**

Preparación de los Datos:

Definir objetivos y Establecer sus métricas





Experto IN	Crea scorecards (i) para el caso de negocio inicial, y (ii) como la base para el seguimiento en la web
Usuario	Configura y despliega los informes de calidad de datos

Peligros en AD

- Privacidad
- Seguridad
- Decisiones basado en datos incompletos
- Decisiones con datos inexactos
- Usando sólo los datos que apoyan nuestras decisiones
- Llegar a la conclusión errónea de los datos: por ejemplo, los precios de las acciones

Definición (Evento, atributo) Sea E el universo de eventos, es decir, el conjunto de todos los posibles identificadores de eventos. Los eventos pueden caracterizarse por diversos atributos. Asumimos los siguientes atributos estándar:

- #actividad (e) es la actividad asociada al evento e.
- #time (e) es la marca de tiempo del evento e.
- #resource (e) es el recurso asociado al evento e.
- #trans (e) es el tipo de transacción asociado al evento e,

Definición Secuencias: son la forma más natural de presentar trazas en un registro de eventos. Dada su importancia, nosotros introducir algunos operadores útiles en las secuencias.

- Una secuencia está representado por una cadena de eventos de longitud dada. $| \sigma |$ denota la longitud de la secuencia, es decir, $| \sigma | = n$.
- dos secuencias se pueden agregar $\sigma 1 \oplus \sigma 2$ dando como resultado una secuencia de longitud $| \sigma 1 | + | \sigma 2 |$.
- una secuencia tiene una cabeza $hd_k(\sigma)$ que consiste de los primeros k elementos (si es posible). $hd_0(\sigma)$ es la secuencia vacía
- Una secuencia tiene una cola $tl_k(\sigma)$ que es la secuencia compuesta de los últimos k elementos. $tl_0(\sigma)$ es la secuencia vacía
- $\sigma \uparrow X$ es la proyección de σ en algún subconjunto $X \subseteq A$,
- Cualquier secuencia ir en conjunto de datos

Desafíos al extraer registros de eventos

Desafío 1: Correlación

Los eventos en un registro de eventos se agrupan por caso. Este simple requisito puede ser bastante desafiante ya que requiere la correlación de eventos, es decir, los eventos deben estar relacionados entre sí. Considere, por ejemplo, datos de eventos dispersos en varias tablas o incluso en varios sistemas. ¿Cómo identificar eventos y sus casos correspondientes?

Desafío 2: marcas de tiempo

Los eventos deben ordenarse por caso. En principio, tal ordenamiento no requiere marcas de tiempo. Sin embargo, cuando se combinan datos de diferentes fuentes, uno normalmente necesita depender de las marcas de tiempo para ordenar los eventos (en orden de aparición). Esto puede ser problemático debido a múltiples relojes y registros demorados, tiempos más largos de medida, etc. Como resultado, es imposible reconstruir el orden de los eventos en un día determinado.

- Una forma de abordar este problema es asumir solo un orden parcial de eventos (es decir, no un orden total) y posteriormente utilizar algoritmos de minería de procesos dedicados para esto.
- Otra forma es "adivinar" el orden según el conocimiento del dominio o los patrones frecuentes a lo largo de los días.

Desafíos al extraer registros de eventos

Desafío 3: instantáneas

Los casos pueden tener una vida útil que se extiende más allá del período registrado, por ejemplo, un caso se inició antes del comienzo del registro de eventos o aún se estaba ejecutando cuando se detuvo la grabación. Por lo tanto, es importante darse cuenta de que los registros de eventos normalmente solo proporcionan una instantánea de un proceso de ejecución más largo. Cuando la duración promedio grabación es corta en comparación con la duración del caso, es mejor eliminar casos incompletos. En muchos casos, las actividades iniciales y finales son conocidas, lo que facilita la filtración del registro de eventos: simplemente elimine todos los casos con "cabeza" o "cola" faltante.

Desafíos al extraer registros de eventos

Desafío 4: Alcance

El cuarto problema es el alcance del registro de eventos. Información de la empresa los sistemas pueden tener miles de tablas con datos relevantes para la empresa. ¿Cómo saber qué tablas incorporar? El conocimiento del dominio es necesario para localizar los datos requeridos y para determinar su alcance.

Desafío 5: granularidad

En muchas aplicaciones, los eventos en el registro de eventos se encuentran en un nivel diferente de granularidad que las actividades relevantes para los usuarios finales. Algunos sistemas producen eventos de bajo nivel que son demasiado detallados. Afortunadamente, hay varios enfoques para preprocesar registros de eventos de bajo nivel.

Calidad de eventos

A nivel de entidad hay tres problemas potenciales:

- **(Falta en el registro)** La entidad existe (o existió) en realidad, pero no fue registrada. Por ejemplo, ocurrió un evento (por ejemplo, tomar una muestra de sangre) pero no fue capturado por el sistema de información.

- **(Desaparecido en la realidad)** La entidad no existe y nunca existió en la realidad, pero fue grabada. Por ejemplo, una cita médica programada nunca se realizó debido a una emergencia, pero de todos modos fue registrada por el sistema de información.

- **(Oculto en el registro)** La entidad se registró y existe (o existió) en la realidad, pero está oculta en un conjunto de datos menos estructurado de mayor tamaño. Por ejemplo, la misma entidad puede aparecer varias veces en el registro de eventos.

Calidad de eventos

clasifica los problemas relacionados con los atributos del evento.

- **(Atributo perdido)** El atributo no se ha registrado para un evento en particular. Por ejemplo, falta la marca de tiempo de un evento.

- **(Atributo incorrecto)** El valor registrado del atributo de evento es incorrecto.

- **(Atributo impreciso)** El valor del atributo de evento es demasiado impreciso. Por ejemplo, el valor de una indicación de fecha y hora es demasiado grueso o la dirección es incompleta.