



Fundamentos de Data Science

Unidad 2: Ciclo de vida de los datos

Semana 6 - Procesamiento de los datos



Unidad 2

Ciclo de vida de los datos

Objetivos

- Reconocer la diferencia entre el procesamiento y el preprocesamiento de los datos.
- Reconocer la importancia del preprocesamiento de los datos dentro del ciclo de vida de los datos.
- Aprender cuales son las principales técnicas aplicadas en el preprocesamiento de los datos.

Procesamiento de los datos

Contenido

1. Procesamiento vs Preprocesamiento de datos
2. Importancia del Preprocesamiento de datos
3. Técnicas de Preprocesamiento de datos



Conclusiones y Preguntas

1. Procesamiento vs. Pre-procesamiento de datos



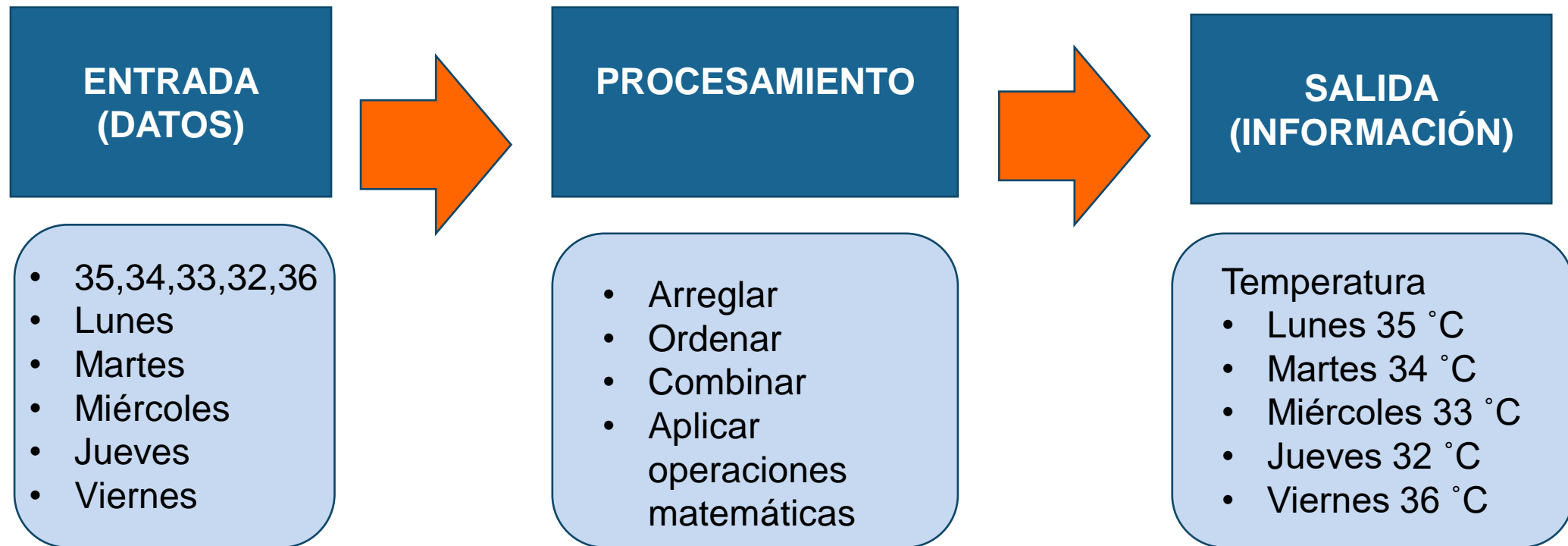
**¿Procesamiento y
Pre-procesamiento de datos
es lo mismo?**

NO

1. Procesamiento vs. Pre-procesamiento de datos

01

Procesamiento de datos



Procesamiento de datos es la conversión de un valor en información útil y deseada

1. Procesamiento vs. Pre-procesamiento de datos

EJEMPLO: CUAL ES LA TEMPERATURA DEL PACIENTE?

DATO



INFORMACION

EL DATO ADQUIERE **SIGNIFICADO**.....

42 grados

PERO UN **SIGNIFICADO** EN CONCRETO

Centígrados

Kelvin

Fahrenheit



Conjunto de DATOS

- Formato (p.e. numérico, categórico, texto, imagen, audio, etc.)

Pero....

- ¿Estos datos estarán completos?
- ¿Serán válidos, confiables?

El DATO

Es un valor respecto de algo que se observa

El **DATO** se convierte en **INFORMACION VERAZ** cuando es capaz de responder a una pregunta concreta y adquiere un significado en concreto

1. Procesamiento vs. Pre-procesamiento de datos

02

Pre-Procesamiento de datos

EN EL MUNDO REAL:

Los conjuntos de datos, en su forma original (datos/hechos brutos), pueden presentar muchas piezas faltantes o estar desordenados.



DEBEMOS:

someterlos a ciertas pasos que garanticen su ***veracidad, completitud y calidad*** para **estar preparados** para un posterior análisis.



Pre-procesando los datos

Intentamos solucionar los problemas que pueden ocurrir durante la recopilación/recolección de datos, antes de procesarlos y convertirlos en información.

1. Procesamiento vs. Pre-procesamiento de datos

02

Pre-Procesamiento de datos

¿Y qué problemas pueden sufrir los datos para que necesiten ser **pre-procesados**?



Los datos adquiridos pueden ser inconsistentes debido a:

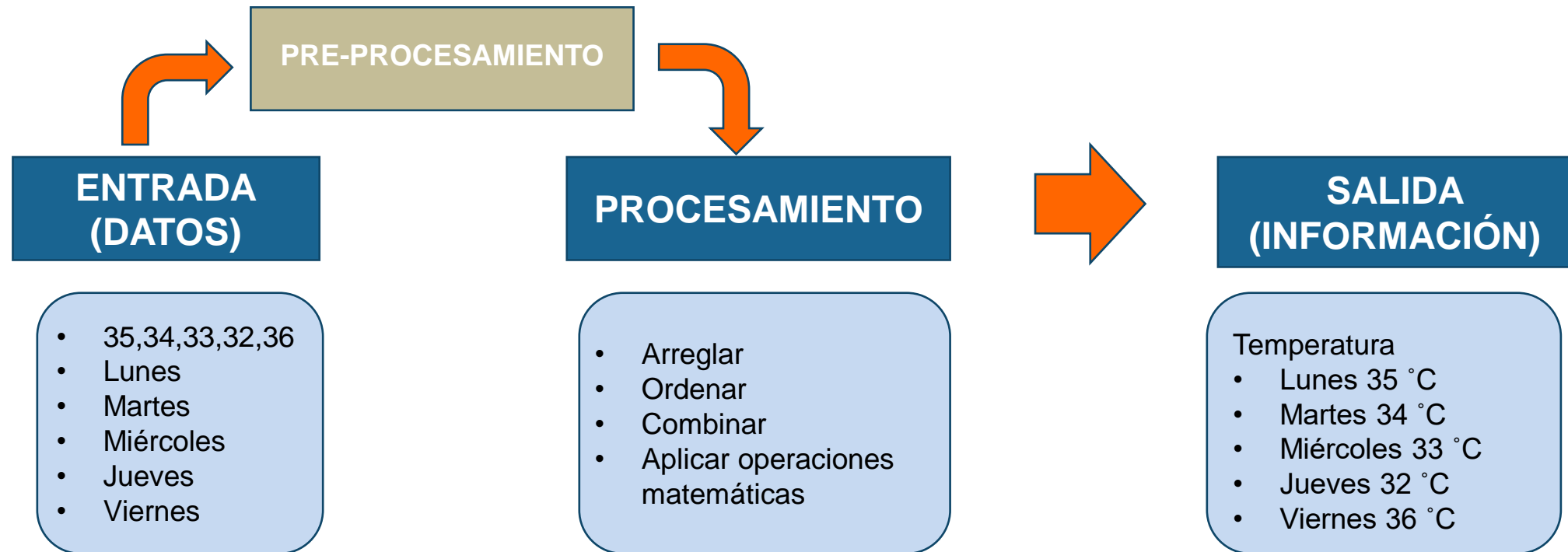
- ☐ Instrumentos defectuosos para la recopilación de datos
- ☐ Errores humanos o informáticos
- ☐ Errores en la transmisión de datos
- ☐ Limitaciones tecnológicas (p. ej., los datos de los sensores llegan a un ritmo más rápido de lo que pueden procesarse)
- ☐ Inconsistencias en las convenciones de nomenclatura o códigos de datos (p. ej., 2/5/2021 podría ser 2 Mayo de 2021 o 5 de febrero de 2021)
- ☐ Duplicación de registros (se recibieron dos veces y debe eliminarse el duplicado)

1. Procesamiento vs. Pre-procesamiento de datos

02

Pre-Procesamiento de datos

El **Pre-procesamiento de datos** prepara los datos (les proporciona veracidad, completitud, calidad) antes de ser procesados y convertidos en información.

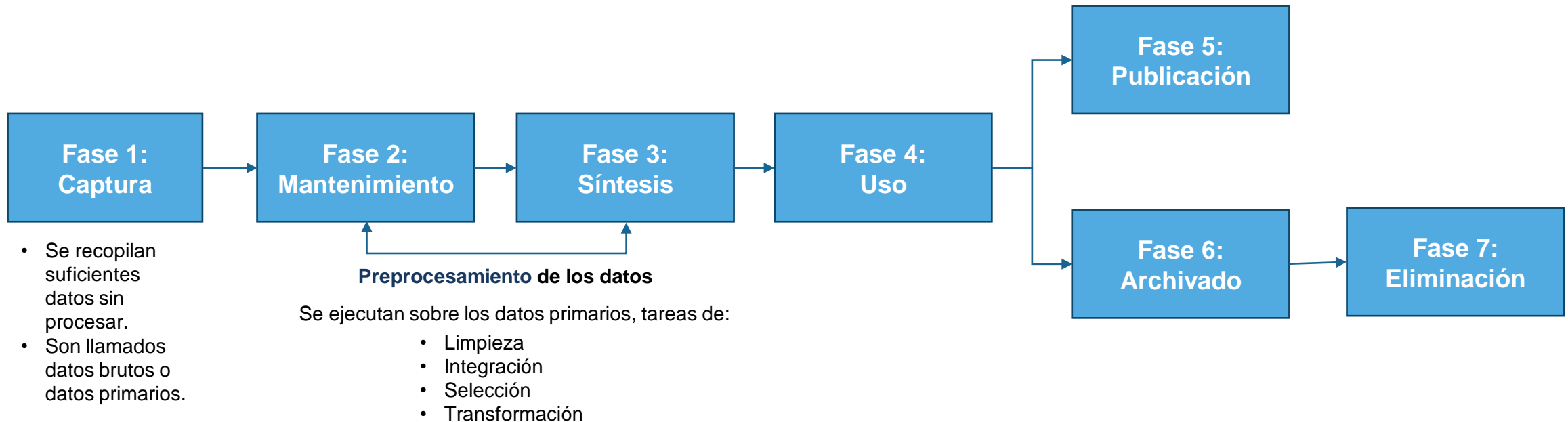


1. Procesamiento vs. Pre-procesamiento de datos

02

Pre-Procesamiento de datos

El Preprocesamiento de los datos dentro del Ciclo de vida del dato



1. Procesamiento vs. Pre-procesamiento de datos

02

Pre-Procesamiento de datos

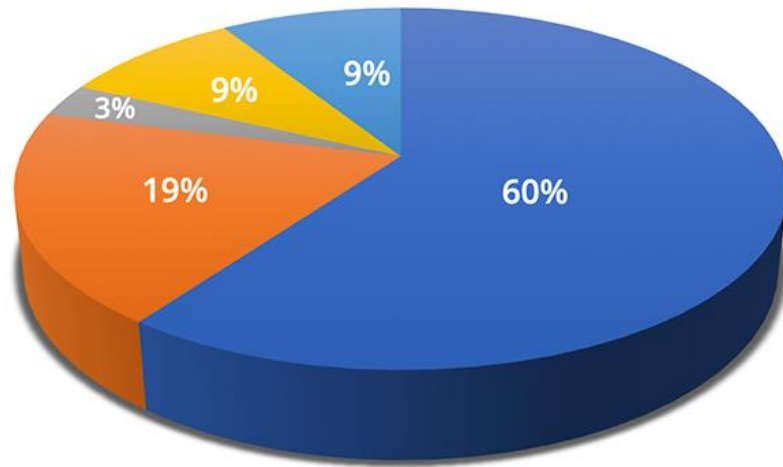
El **pre-procesamiento de datos** es una técnica de **minería de datos** que implica preparar los datos a un formato comprensible y de calidad.

Pre-Procesamiento de datos



2. Importancia del Pre-procesamiento de datos

¿A qué dedica el tiempo un científico de datos?



■ Preprocesamiento de datos ■ Obtener datos ■ Construir modelos ■ Explorar datos ■ Otros

Los científicos de datos invertimos incluso más del 60% del tiempo en aplicar técnicas de pre-procesamiento de datos.

¿Pero, por qué?



2. Importancia del Pre-procesamiento de datos

Porque debemos asegurarnos de trabajar con **DATOS DE CALIDAD**



Debemos de ser capaces de asegurar que los datos con los que trabajamos son:

☐ **FIABLES:**

La **confiabilidad de los datos** (números, eventos y datos históricos) se adquiere cuando están actualizados, Consolidados y son correctos.

☐ **COMPLETOS:**

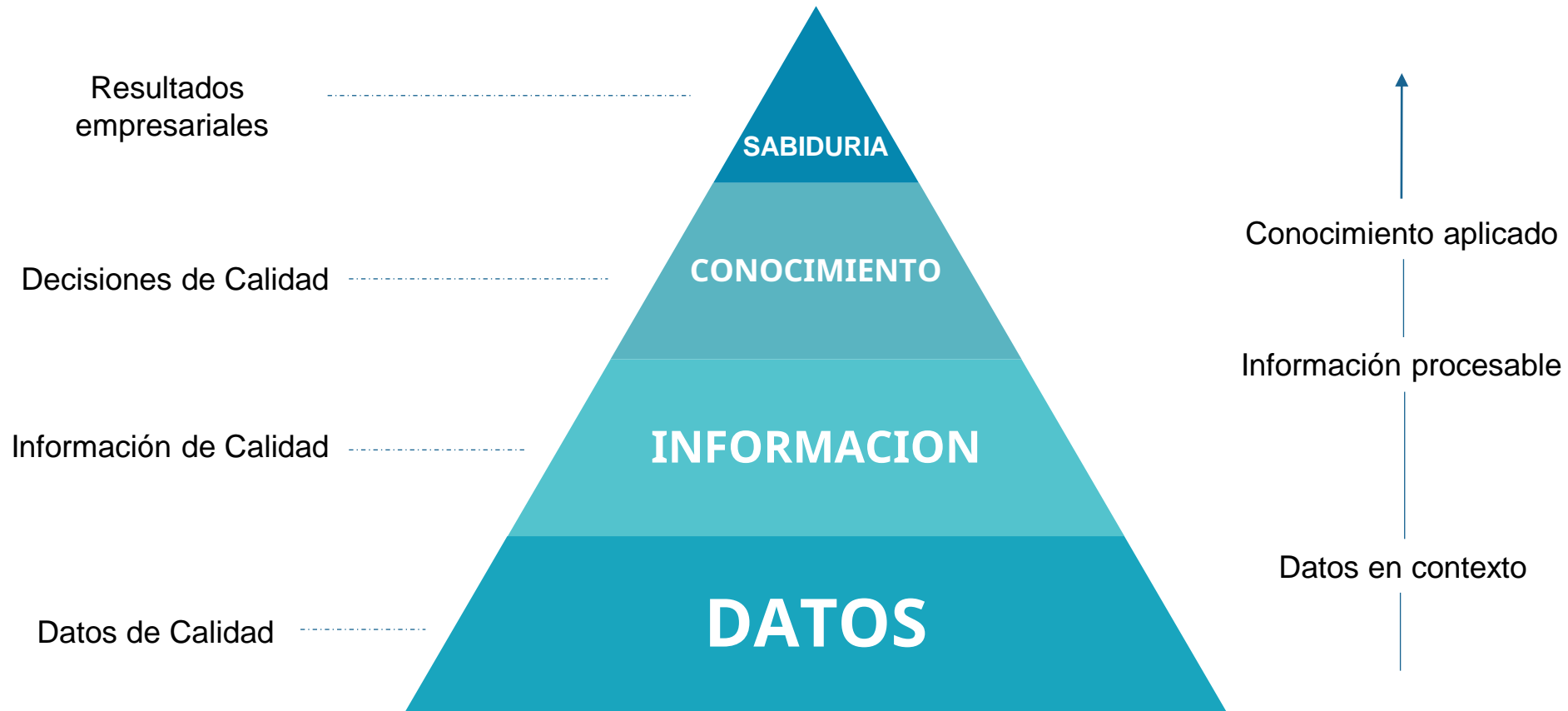
La **completitud de los datos** nos asegura que los atributos de interés están disponibles, que se registraron todos los datos necesarios o si se perdieron ya fueron recuperados.

☐ **CONSISTENTES:**

La coherencia de los datos nos capacita para poder compararlos, segmentarlos, filtrarlos y categorizarlos.

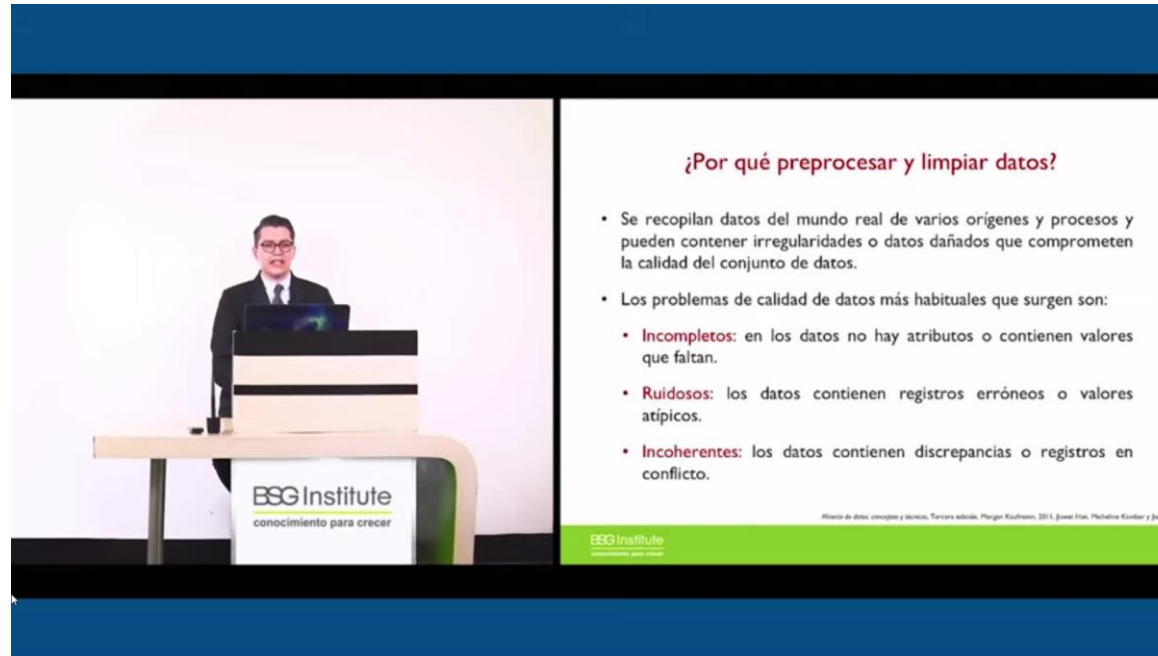
2. Importancia del Pre-procesamiento de datos

Y porque las **decisiones de calidad** deben basarse en **datos de calidad**



2. Importancia del Pre-procesamiento de datos

¿Por qué debemos pre-procesar los datos?



¿Por qué preprocesar y limpiar datos?

- Se recopilan datos del mundo real de varios orígenes y procesos y pueden contener irregularidades o datos dañados que comprometen la calidad del conjunto de datos.
- Los problemas de calidad de datos más habituales que surgen son:
 - **Incompletos:** en los datos no hay atributos o contienen valores que faltan.
 - **Ruidosos:** los datos contienen registros erróneos o valores atípicos.
 - **Incoherentes:** los datos contienen discrepancias o registros en conflicto.

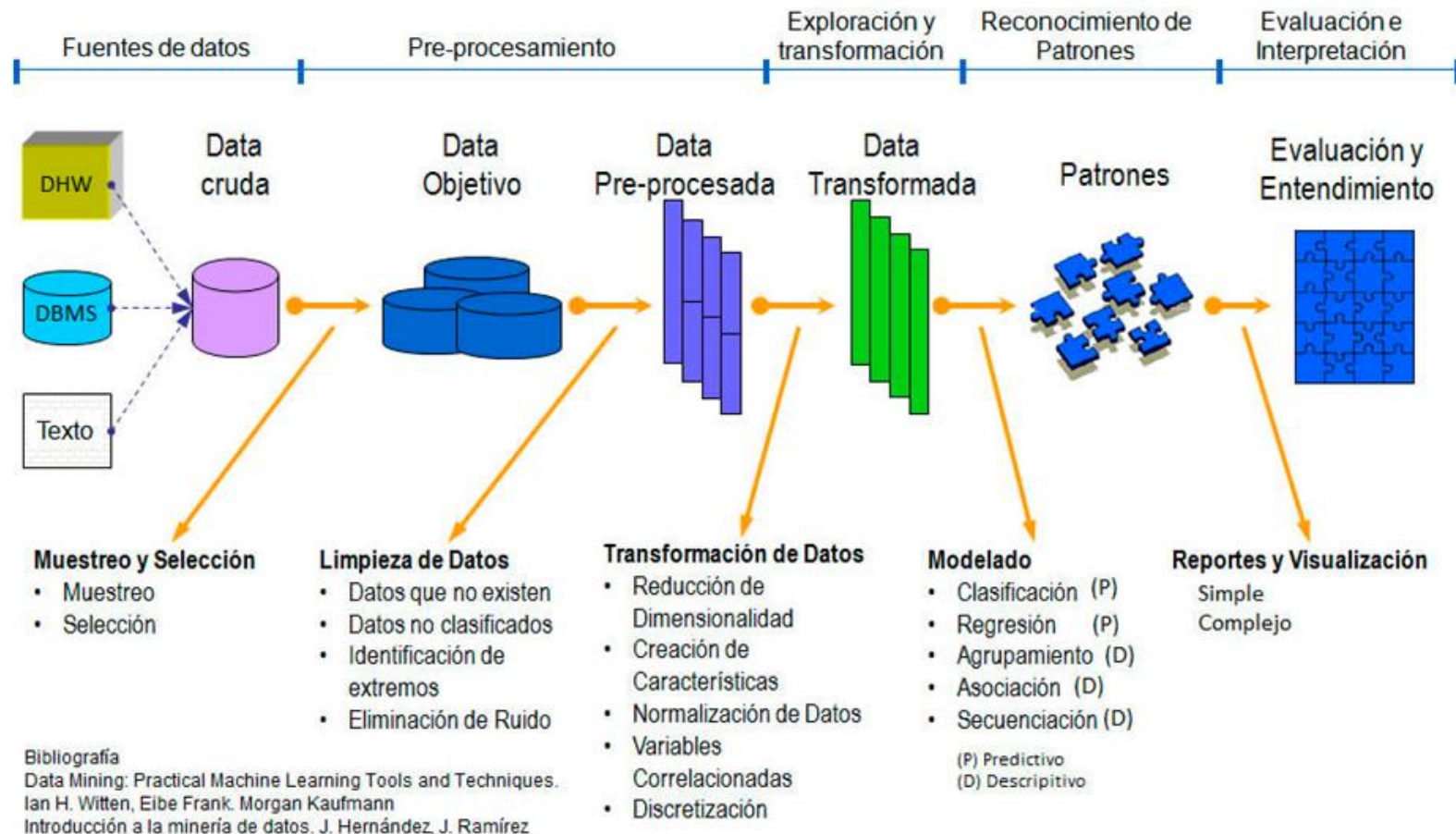
Algoritmo de datos: conceptos y técnicas, Tercera edición, Morgan Kaufmann, 2011, Joost Van, Michaela Kunder y Jan Pei

BSG Institute
conocimiento para crecer

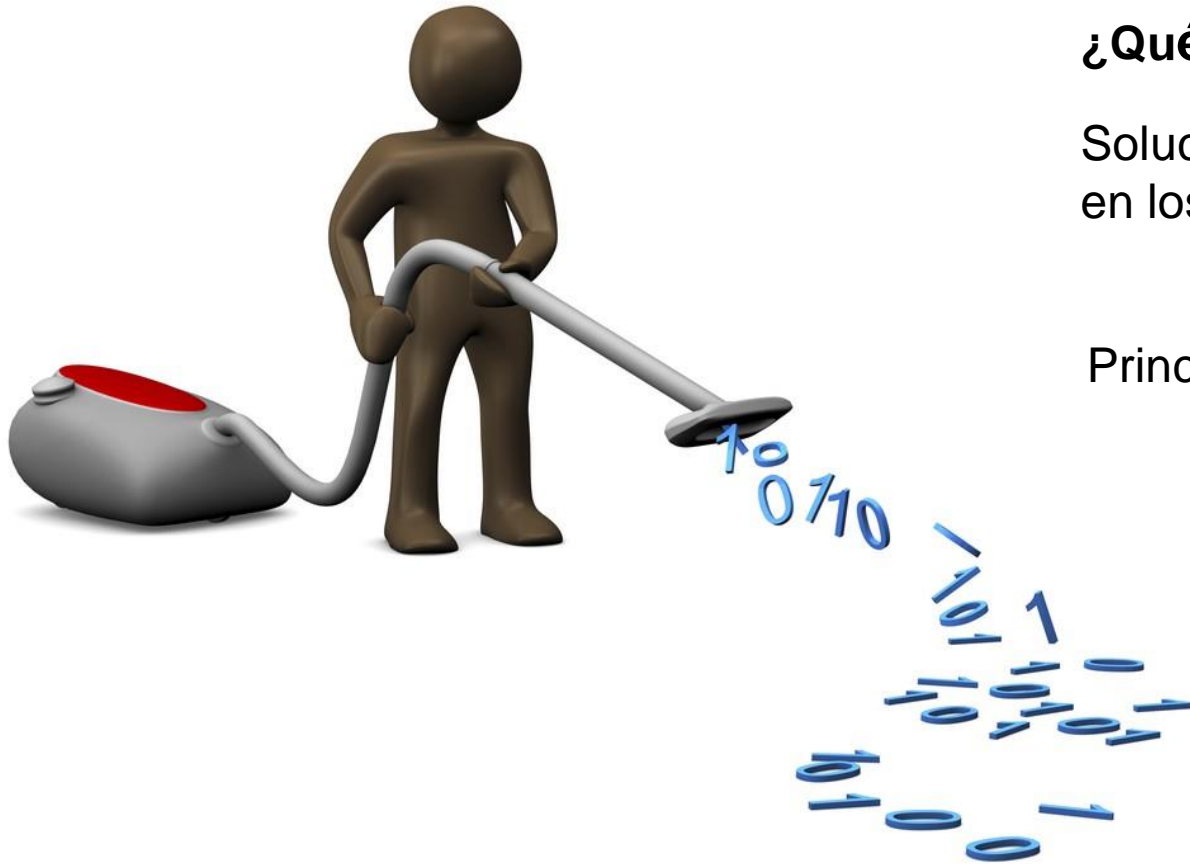
Video: https://youtu.be/xVWTNb_Moro (Duración: 4:30 min)

3. Técnicas de Pre-procesamiento de datos

MINERIA DE DATOS: TECNICAS DE PRE-PROCESAMIENTO DE LOS DATOS



3. Técnicas de Pre-procesamiento de datos



Las Técnicas de pre-procesamiento de datos...
¿Qué hacen?

Solucionan los defectos más comunes encontrados en los conjuntos de datos.

Principales técnicas de pre-procesamiento de datos:

1. Limpieza de datos (depuración de datos)
2. Integración y transformación de datos
3. Reducción de datos

3. Técnicas de Pre-procesamiento de datos

RECORDEMOS LOS TIPOS DE DATOS: Tres estructuras de datos diferentes



01

Limpieza de datos

02

Integración y transformación de datos

03

Reducción de datos

3. Técnicas de Pre-procesamiento de datos

01

Limpieza de datos (depuración de datos)

DATOS TRADICIONALES

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Annotations:

- Missing values: Row 2, City column.
- Invalid values: Row 5, Gender column (A).
- Misfielded values: Row 7, City column (Italy).
- Misspellings: Row 8, Country column (Portugal); Row 10, Country column (Ytali).
- Uniqueness: Row 5, Id column (555); Row 6, Id column (555).
- Formats: Row 6, Birthday column (1983-12-01).
- Attribute dependencies: Row 9, #Students column (5).

Problemas a resolver:

- ☐ Completar los valores faltantes
- ☐ Identificar valores atípicos (outliers)
- ☐ Corregir los datos inconsistentes (ortográficos, formato, etc.)
- ☐ Eliminar registros duplicados

¿Cómo manejar los datos faltantes?

- Ignorar la tupla: generalmente se hace cuando falta la etiqueta de la clase (asumiendo las tareas en la clasificación); no es efectivo cuando el porcentaje de valores perdidos por atributo varía considerablemente.
- Completar el valor faltante manualmente: tedioso + inviable
- Utilizar una constante global para completar el valor que falta: por ejemplo, "desconocido", ¿una nueva clase?
- Utilizar el atributo "media" para completar el valor perdido
- Utilizar el atributo "media" para todas las muestras que pertenecen a la misma clase para completar el valor faltante: más inteligente
- Utilizar el valor más probable para completar el valor faltante: basado en inferencias Fórmula bayesiana o árbol de decisión.

3. Técnicas de Pre-procesamiento de datos

02

Integración y transformación de datos

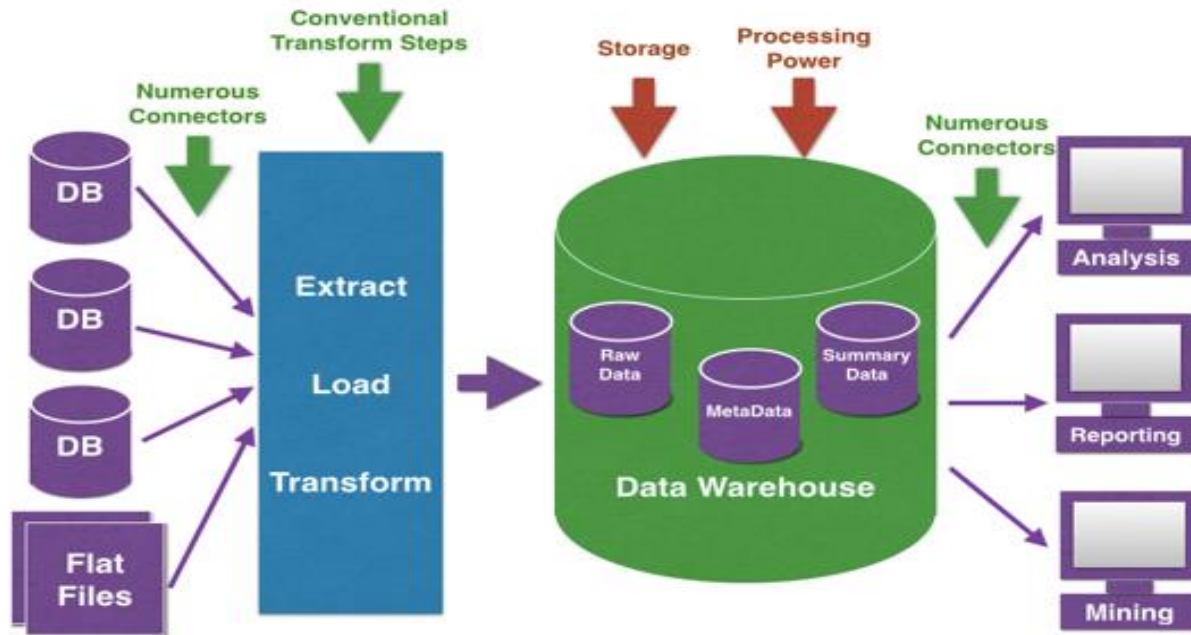


Figure 1 - Traditional Data Integration

DATOS TRADICIONALES -Problemas a resolver:

- ☐ Identificación de entidades (misma entidad con distintos identificadores)
- ☐ Conflictos de valor de datos (valores del mismo atributo de diferentes fuentes son diferentes)
- ☐ Diferentes escalas, por ejemplo, unidades métricas frente a unidades británicas (pulgadas frente a centímetros)
- ☐ Datos redundantes (ocurren a menudo cuando se integran múltiples bases de datos)

3. Técnicas de Pre-procesamiento de datos

02

Integración y transformación de datos

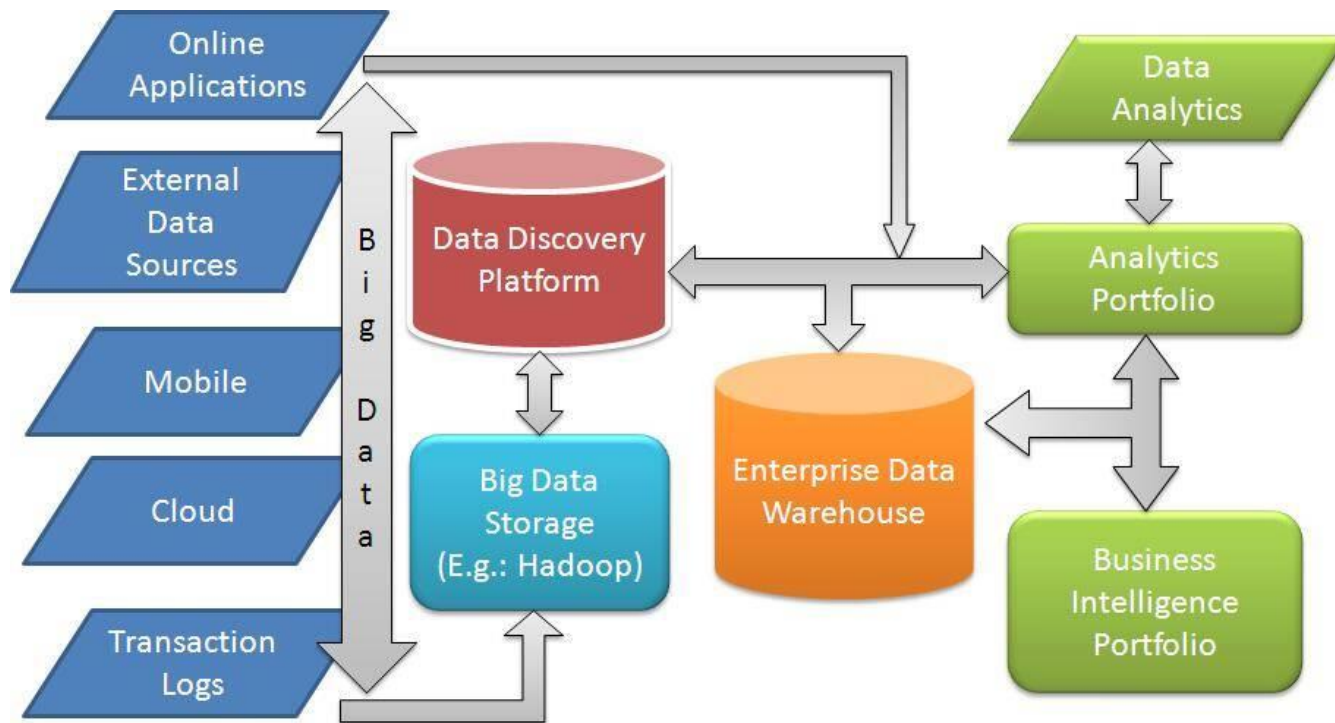


Figura 2. Integración de datos tradicionales y macrodatos

MACRODATOS - Problemas a resolver:

Pueden tener los mismos problemas que los datos tradicionales (pero en mayor volumen, variedad y velocidad...).

Y adicionalmente:

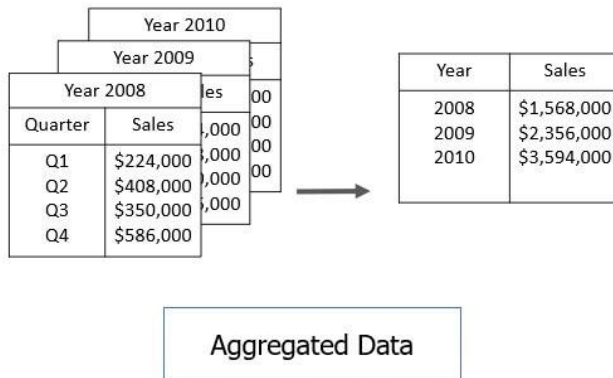
- ❑ Los macrodatos tienen más tipos de datos, por ello, existe una variedad más amplia de métodos de limpieza de datos.
 - Datos de texto
 - Datos de imagen digital
 - Datos de video digital
 - Datos de audio digital
- ❑ Existen técnicas que verifican si una imagen digital está lista para su procesamiento.
- ❑ Existen enfoques específicos que garantizan que la calidad de audio de su archivo sea la adecuada para continuar.

3. Técnicas de Pre-procesamiento de datos

02

Integración y transformación de datos

Estrategias más utilizadas



Agregación

- Es el proceso de recopilar datos y presentarlos en un formato resumido. Los datos se pueden recopilar de múltiples fuentes de datos con la intención de combinar estas fuentes de datos en un resumen para el análisis de datos.

Less generalized				More generalized		
ID	Age	Diagnosis	Anonymous	Age	Diagnosis	Anonymous
0	[20-39]	Colon cancer	Yes	[20-79]	Colon cancer	No
1	[20-39]	Stroke		[20-79]	Stroke	
2	[20-39]	Colon cancer		[20-79]	Colon cancer	
3	[40-59]	Colon cancer	Yes	[20-79]	Colon cancer	
4	[40-59]	Stroke		[20-79]	Stroke	
5	[60-79]	Stroke	No	[20-79]	Stroke	
6	[60-79]	Stroke		[20-79]	Stroke	
7	[60-79]	Stroke		[20-79]	Stroke	
8	[60-79]	Stroke		[20-79]	Stroke	
9	[60-79]	Stroke		[20-79]	Stroke	
10	[60-79]	Stroke		[20-79]	Stroke	
11	[60-79]	Stroke		[20-79]	Stroke	
12	[60-79]	Stroke		[20-79]	Stroke	
13	[60-79]	Stroke		[20-79]	Stroke	
14	[60-79]	Stroke		[20-79]	Stroke	

Generalización

- Es el proceso de resumir datos mediante la sustitución de valores de nivel relativamente bajo con conceptos de nivel superior. Es una forma de minería de datos descriptiva.

Cambie el rango de:

$[-\infty, +\infty]$ a $[-1, +1]$

$\{-13, -6, -3, 100\}$ $\{-0.13, -0.06, -0.03, 0, 1, 1\}$

Normalización

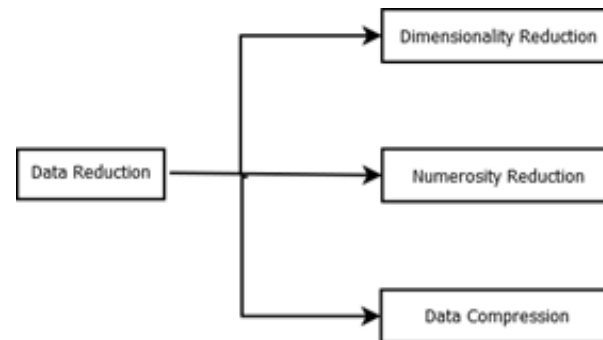
- Es el proceso de cambiar los valores de las columnas numéricas en el conjunto de datos a una escala común, sin distorsionar las diferencias en los rangos de valores.
- Se realiza para caer dentro de un rango pequeño y específico.
- Por lo general, la normalización significa cambiar la escala de los valores en un rango de $[0,1]$.

3. Técnicas de Pre-procesamiento de datos

03

Reducción de datos

Problema a resolver: El proceso análisis/minería de datos complejos puede tardar mucho en ejecutarse en un conjunto de datos completo.



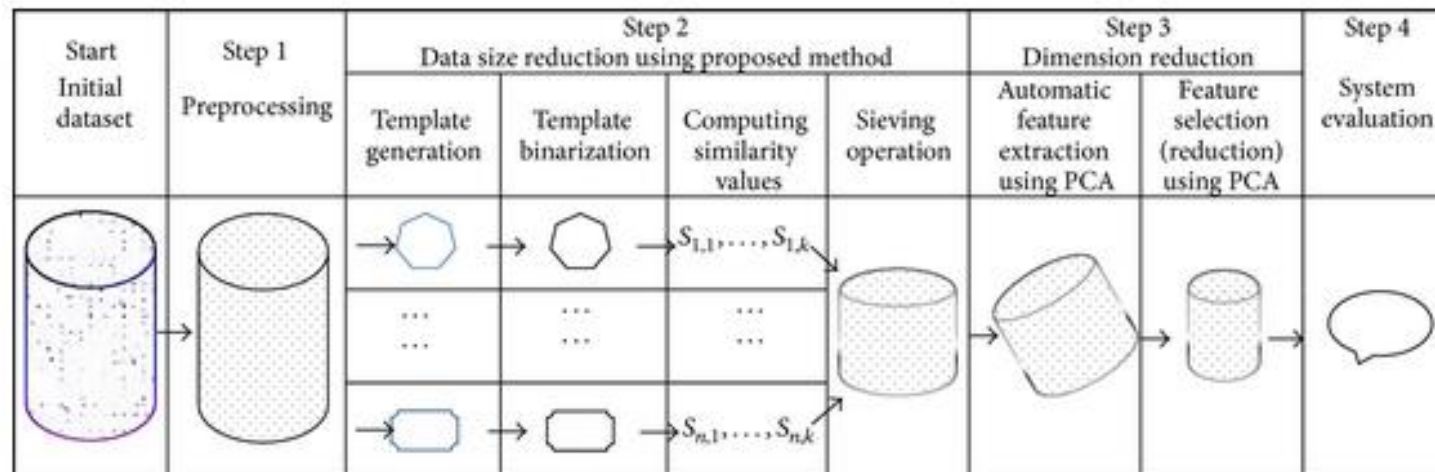
Reducción de datos: Obtenemos una representación reducida del conjunto de datos que es mucho más pequeño en volumen, pero sin embargo produce los mismos (o casi iguales) resultados analíticos

3. Técnicas de Pre-procesamiento de datos

03

Reducción de datos

Problema a resolver: El análisis / minería de datos complejos puede tardar mucho en ejecutarse en el conjunto de datos completo.



Fuente: <https://www.hindawi.com/journals/mpe/2014/537428/>

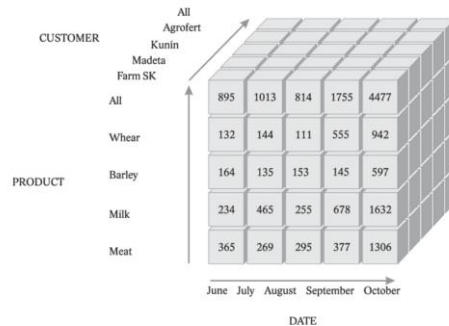
Reducción de datos: Obtenemos una representación reducida del conjunto de datos que es mucho más pequeño en volumen, pero sin embargo produce los mismos (o casi iguales) resultados analíticos.

3. Técnicas de Pre-procesamiento de datos

03

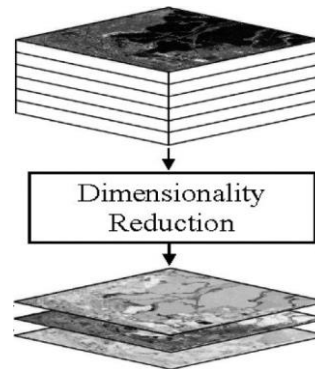
Reducción de datos

Estrategias de Reducción de datos



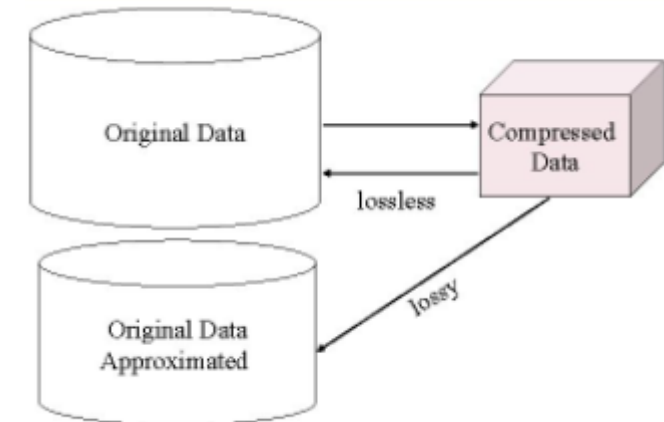
Agregación de Cubos de Datos

- Un cubo de datos es un modelo de datos multidimensional que almacena los datos optimizados, resumidos o agregados, lo que facilita las herramientas OLAP para un análisis rápido y sencillo.
- El cubo de datos almacena los datos pre-calculados y facilita el procesamiento analítico en línea.



Reducción de Dimensionalidad

Selección de características (es decir, selección de subconjuntos de atributos) de modo que la distribución de probabilidad de diferentes clases dados los valores de esas características sea lo más cercana posible a la distribución original.



Compresión de datos

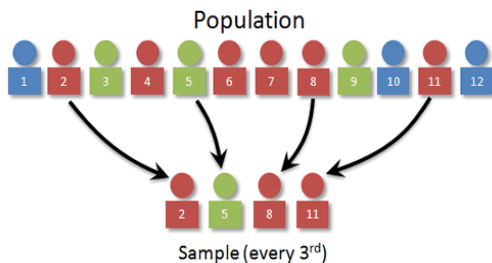
Datos comprimidos:
De los datos originales sin pérdidas
Obtenemos datos originales aproximados (menor volumen).

3. Técnicas de Pre-procesamiento de datos

03

Reducción de datos

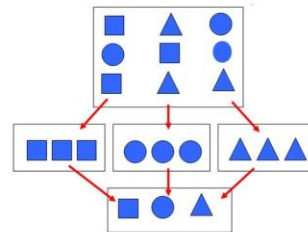
Estrategias de Reducción de datos



Reducción de Numerosidad

Obtenemos una muestra (sampling) pequeña que represente al conjunto de datos

Stratified Random Sampling



CID	Age	Gender	Married	Salary
c1	[20,25)	Male	No	[2000, 2900)
c2	[25,30)	Female	No	[2000, 2900)
c3	[20,25)	Male	No	[2900, 3800)
c4	[25,30)	Female	Yes	[2000, 2900)
c5	[30,35)	Male	Yes	[3800, 4700]
c6	[35,40]	Male	Yes	[2900, 3800)

Discretización

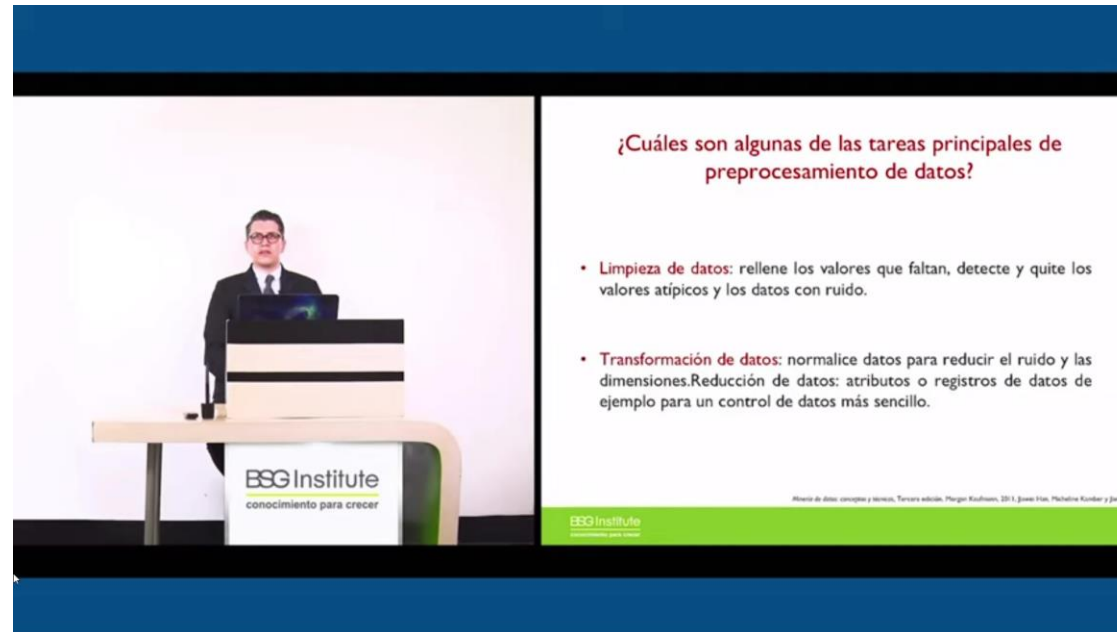
Divide el rango de un atributo continuo en intervalos

¿Por qué?

- Algunos algoritmos de clasificación solo aceptan atributos categóricos.
- Reduce el tamaño de los datos
- Prepara los datos para análisis adicionales

3. Técnicas de Pre-procesamiento de datos

Tareas o técnicas principales de pre-procesamiento de datos



¿Cuáles son algunas de las tareas principales de preprocesamiento de datos?

- **Limpieza de datos:** rellene los valores que faltan, detecte y quite los valores atípicos y los datos con ruido.
- **Transformación de datos:** normalice datos para reducir el ruido y las dimensiones. Reducción de datos: atributos o registros de datos de ejemplo para un control de datos más sencillo.

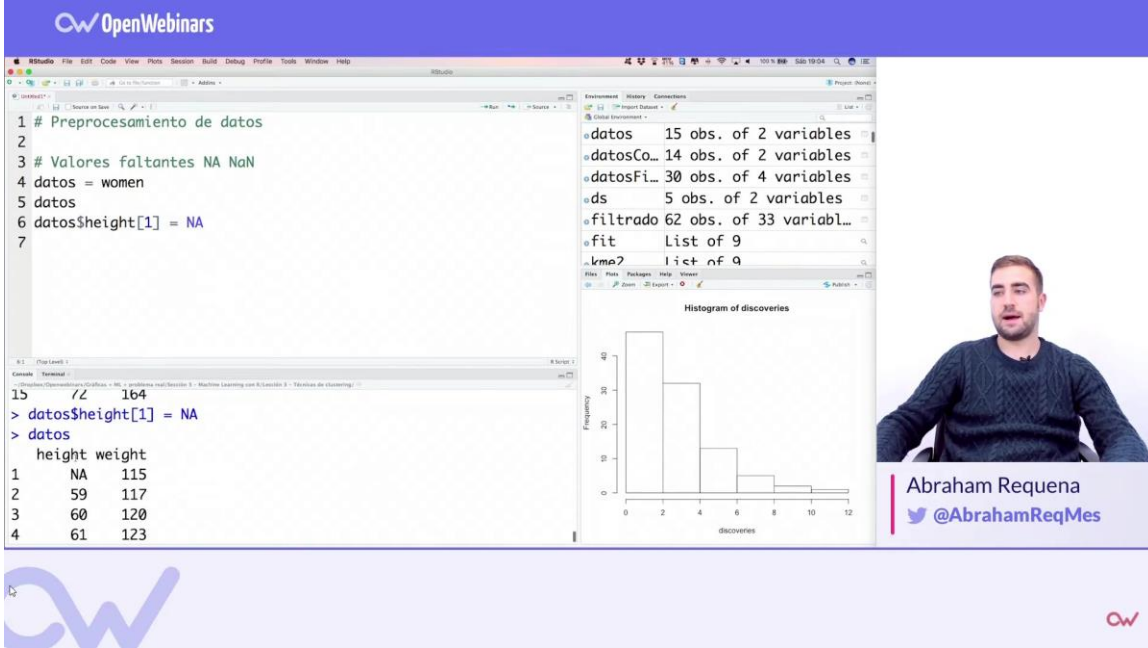
Minicurso de datos: conceptos y técnicas. Tercera edición. Plurigen Knowledge, 2013. Juan Vico, Michael Kuebler y Juan Pe

BSG Institute
conocimiento para crecer

Video: <https://youtu.be/edbsA5DugVo> (Duración: 4 min)

3. Técnicas de Pre-procesamiento de datos

Pre-procesamiento: Limpieza de datos en R



The screenshot displays an RStudio interface with a script editor on the left containing R code for data preprocessing. The code includes comments in Spanish and a command to set a value to NA. The console on the bottom left shows the execution of the code, resulting in a data frame with 4 rows and 2 columns: height and weight. The environment pane on the right lists several objects, including 'datos', 'datosCo...', 'datosFi...', 'ds', 'filtrado', 'fit', and 'lmpa?'. A histogram titled 'Histogram of discoveries' is also visible in the bottom right of the environment pane. A video inset in the bottom right corner shows a man, Abraham Requena, with his Twitter handle @AbrahamReqMes.

```
1 # Preprocesamiento de datos
2
3 # Valores faltantes NA NaN
4 datos = women
5 datos
6 datos$height[1] = NA
7
```

```
15 72 164
> datos$height[1] = NA
> datos
  height weight
1     NA   115
2     59   117
3     60   120
4     61   123
```

Environment: Global Environment

- datos: 15 obs. of 2 variables
- datosCo...: 14 obs. of 2 variables
- datosFi...: 30 obs. of 4 variables
- ds: 5 obs. of 2 variables
- filtrado: 62 obs. of 33 variables
- fit: List of 9
- lmpa?: List of 9

Histogram of discoveries

Abraham Requena
@AbrahamReqMes

Video: <https://youtu.be/JFMtN5OYkxA> (Duración: 5:50 min)

Conclusiones

1. El **Procesamiento de datos** es la conversión de un valor o dato en información útil y deseada.
2. **Pre-procesamiento de los datos**, se denomina a una serie de tareas o técnicas ejecutadas previamente al procesamiento de los datos.
3. Estas tareas de Pre-procesamiento se aplican sobre los datos brutos (también llamados " hechos sin procesar " o " datos primarios ") porque en ese estado, los datos no pueden ser analizados de inmediato.
4. Los datos adquiridos pueden ser inconsistentes debido a:
 - El uso de instrumentos defectuosos para la recolección de datos.
 - Errores humanos o informáticos
 - Errores en la transmisión de datos
 - Limitaciones tecnológicas (p. ej., los datos de los sensores llegan a un ritmo más rápido de lo que pueden procesarse)
 - Inconsistencias en las convenciones de nomenclatura o códigos de datos (p. ej., 2/5/2021 podría ser 2 Mayo de 2021 o 5 de febrero de 2021)
 - Duplicación de registros (se recibieron dos veces y debe eliminarse alguno de ellos)
5. Las tareas de Pre-procesamiento de datos convertirán los datos sin procesar a un formato que es más comprensible y útil para su procesamiento posterior. Algunas de las técnicas/tareas de pre-procesamiento (preparación de los datos) son:
 - Limpieza de datos
 - Integración y transformación de datos
 - Reducción de datos
6. Las tareas de **Pre-procesamiento de datos** forman parte de la **Fase 2: Mantenimiento del dato** y **Fase 3: Síntesis del dato**, dentro del ciclo de vida del dato. Dichas tareas se ejecutan previamente a la Fase 4: Uso del dato.
7. La técnica de Limpieza de datos, busca resolver los siguientes problemas:
 - Completar los valores faltantes
 - Identificar valores atípicos
 - Corregir los datos inconsistentes (ortográficos, formato, etc.)
 - Eliminar registros duplicados
8. Las técnicas de Integración y transformación de datos, buscan resolver los siguientes problemas:
 - Identificación de entidades (misma entidad con distintos identificadores)
 - Conflictos de valor de datos (valores del mismo atributo de diferentes fuentes son diferentes)
 - Diferentes escalas, por ejemplo, unidades métricas frente a unidades británicas (pulgadas frente a centímetros)
 - Datos redundantes (ocurren a menudo cuando se integran múltiples bases de datos)

Conclusiones

9. Las estrategias mas utilizadas para la Integración y Transformación de datos son:
 - Agregación
 - Generalización
 - Normalización
10. La técnica de **Reducción de datos**, tiene por objetivo obtener una representación reducida del conjunto de datos que es mucho más pequeño en volumen, pero sin embargo produce los mismos (o casi iguales) resultados analíticos.
11. La **Reducción de datos** busca aliviar o reducir el tiempo/costo del proceso de análisis/minería de datos complejos, dado que puede tardar mucho más si este se ejecuta en un conjunto de datos completo. Se pueden aplicar varios métodos de reducción de datos, entre ellos, los principales son:
 - Agregación de cubos de datos
 - Reducción de dimensionalidad
 - Compresión de datos
 - Reducción en numerosidad
 - Discretización de datos



PREGUNTAS

Dudas y opiniones