

# Aprendizaje de Máquina

## Pre-procesamiento de datos

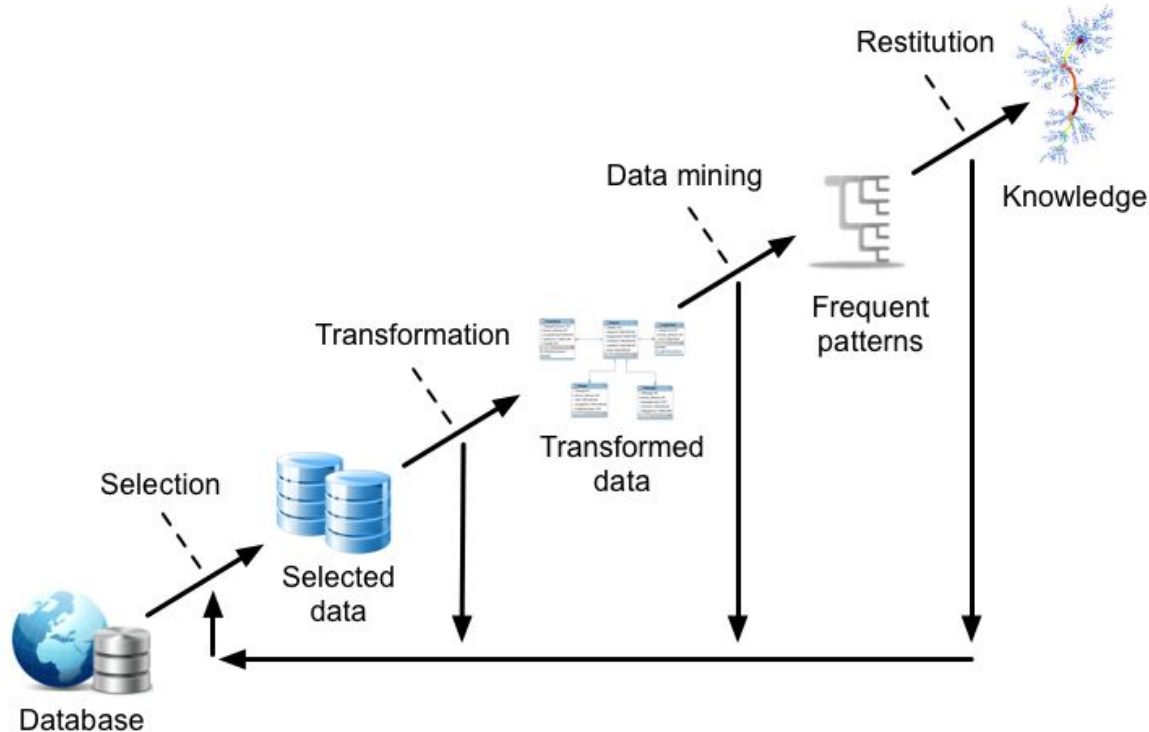
**César A. Beltrán Castañón**

Pontificia Universidad Católica del Perú

# Agenda

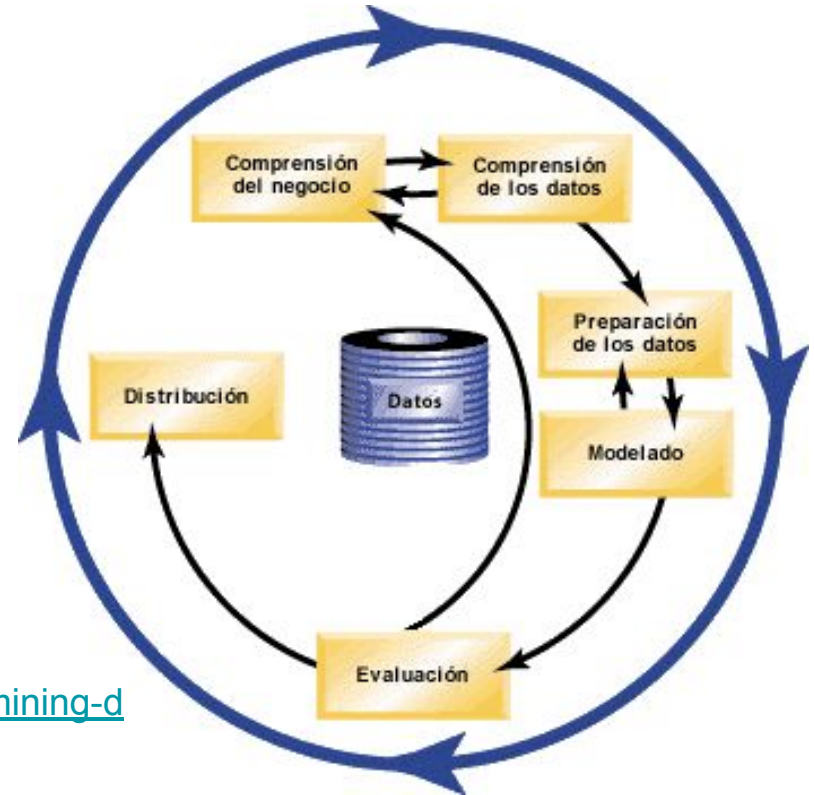
1. **Proceso de Minería de Datos**
2. Acerca de los datos
3. Pre-procesamiento de datos
  - a. Problemas con los datos
  - b. Tareas de Pre-procesamiento
4. Herramientas para el análisis de tratamiento de datos

# Proceso KDD - Knowledge Discovery on Databases



# Modelo CRISP-DM

Cross-Industry Standard Process for Data Mining



<https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>

# Agenda

1. Proceso de Minería de Datos
2. **Acerca de los datos**
3. Pre-procesamiento de datos
  - a. Problemas con los datos
  - b. Tareas de Pre-procesamiento
4. Herramientas para el análisis de tratamiento de datos

# ¿Qué son los datos?

[Wikipedia] Data is a **set of values** of qualitative or quantitative variables;  
restated, pieces of data are individual pieces of **information**

# Acercas de los datos

The diagram illustrates a data table with three key components highlighted by blue brackets:

- Atributos:** A bracket above the table columns, indicating the features or attributes.
- Objetos:** A bracket to the left of the table rows, indicating the individual data instances or objects.
- Clase:** A bracket below the 'Alerta' column, indicating the target variable or class.

Distrito	Fecha	Temperatura	Humedad	Viento	Alerta
San Miguel	14/03/12	14.4	68	57	Si
San Miguel	15/03/12	18.4	60		No
Pueblo Libre	14/03/12	20.3	72	45	Si
Pueblo Libre	01/04/12	15.6	68	11	No
Comas	18/04/12	28.0	71		No

# Acercas de los datos

- Los datos permiten representar una colección de objetos y los atributos que los describen
- Un atributo es una propiedad o característica de un objeto
  - e.g., color de ojos de una persona, la temperatura, etc.
  - Los atributos son también conocidos como variables, campos o características
- La colección de atributos describen un objeto
  - Los objetos son conocidos como registros, puntos, casos, entidades, individuos o instancias
- Los objetos pueden o no pertenecer a una clase
  - Ciertas características de un objeto pueden determinar la clase a la cual ellos pertenecen
  - Las clases pueden variar de dos (binaria) a más (multi-clase)



# Tipos de atributos

- **Nominal:** Representan **categorías**, estados o "nombre de cosas"
  - e.g., días de la semana = {domingo, lunes, martes, miércoles, jueves, viernes, sábado}
- Otros ejemplos: estado marital, ocupación, DNI, color de ojos, etc.
- **Ordinal:** Valores que implican un **orden** (ranking)
  - e.g., talla = {pequeño, mediano, grande}
  - e.g., evaluación de la aceptación de las "Papitas Lay's" (escala del 1 -10)
  - La magnitud entre valores sucesivos no es conocida
- **Cardinal:** Representan una **cantidad**
  - e.g., el peso, un salario, el ángulo formado por dos segmentos

# Tipos de atributos

- **Intervalos:** medidas sobre una escala de unidades de igual tamaño
  - e.g., la temperatura en °C o °F, fechas en el calendario (semanas), etc.
  - No existe el valor "0" (true zero-point), e.g., 0°C no es cero absoluto, existen otros valores debajo de este
- **Proporción (ratio):** valores que tienen un **orden** de magnitud más grande que la unidad de medida
  - e.g., la temperatura en °K, la talla, cantidades monetarias
  - e.g., el peso (10 Kg es dos veces 5 Kg)

# Tipos de atributos

- **Binarios**

- Son atributos **nominales** con solo **dos estados**, e.g., 0 y 1
- Binarios **simétricos**: ambos valores tienen la **misma importancia**, e.g., sexo = {masculino, femenino}
- Binarios **asimétricos**: no simétricos :-). e.g., test médicos {positivo, negativo}
- Por convención, asignamos 1 al más importante (e.g., VIH positivo)

# Atributos Discretos vs Continuos

- Atributos **discretos (categóricos)**
  - Posee solamente **valores finitos** o contablemente finitos, e.g., números telefónicos, letras en un documento, profesiones, etc.
  - Algunas veces es representado por un valor **entero**
  - Nota: los atributos binarios son un tipo especial de atributos discretos
- Atributos **continuos**
  - Tienen **números reales** como valores de los atributos, e.g., la temperatura, el peso, la talla, etc.
  - Practicamente, pueden solamente ser medidos y representados por un **número finito** de dígitos
  - Son típicamente representados por variables de **punto flotante**.

# Ejercicios

1. Investigadores de la PUCP se interesan en el **estudio de la obesidad** en jóvenes de tres colegios de Lima. De dos ejemplos de atributos binarios (con sus posibles valores), dos de atributos discretos y dos de atributos continuos que pueden ser utilizados en este estudio
2. Un cuestionario está asociado al **estudio de personas** y se toman en cuenta los siguientes atributos: a) la edad, b) el sexo, c) la profesión, d) la talla, y; e) el número de hijos. Indique los tipos de datos utilizados para cada atributo
3. Un instructor **anota el orden** en el cual sus estudiantes terminan sus ejercicios. El primero ya termino, el segundo, ... ¿Que tipo de atributos se están utilizando?
4. En un estudio sobre la percepción de expresiones faciales, se pretende **determinar la polaridad** en la expresión de los individuos estudiados. ¿Que tipo de atributo se utilizará y qué valores posibles puede tener este atributo?
5. Se pretende estudiar el impacto de la **temperatura** en la aparición de una **epidemia de malaria**. Los datos se recogen diariamente por un año. ¿Como agrupará los datos de manera a que sean binarios? por ejemplo, los agrupo por trimestres pero el atributo no es binario, sino nominal.

# Respuestas

1. Binarios: sexo (masculino, femenino), fumador (si, no), practica un deporte (si, no), etc.  
Discretos: número de comidas al día, número de horas pasadas frente al televisor, etc.  
Continuos: peso, talla, etc.
2. Muchas formas de responder esta pregunta. Discreto a), c) y d), continuo d), binario b)
3. Ordinal
4. Nominal, con valores positivo, neutro y negativo
5. periodo={epidémico, no epidémico}

# Agenda

1. Proceso de Minería de Datos
2. Acerca de los datos
3. **Pre-procesamiento de datos**
  - a. Problemas con los datos
  - b. Tareas de Pre-procesamiento
4. Herramientas para el análisis de tratamiento de datos

# ¿Por qué pre-procesar los datos?

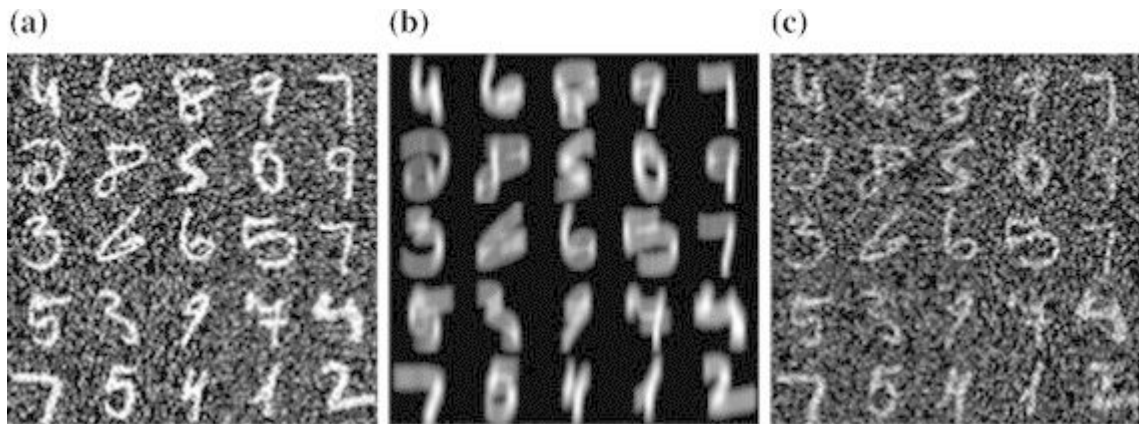
- Los datos en el **mundo real** son “**sucios**”
  - **Son ruidosos:** contiene errores
  - **Contiene valores atípicos:** valores muy diferentes a la media
  - **Son incompletos:** carentes de valores en los atributos, faltan ciertos atributos de interés, o contienen sólo datos agregados
  - **Son inconsistentes:** contiene discrepancias en códigos o nombres
- Datos sin **calidad**, resultados sin calidad!
  - **Decisiones** de calidad deben basarse en datos de calidad
  - Para la minería de datos se necesita la **integración coherente** de datos de calidad

El pre-procesamiento de datos toma generalmente el **60% de todo el proceso KDD**



# El ruido

El ruido está relacionado a la modificación de los datos originales

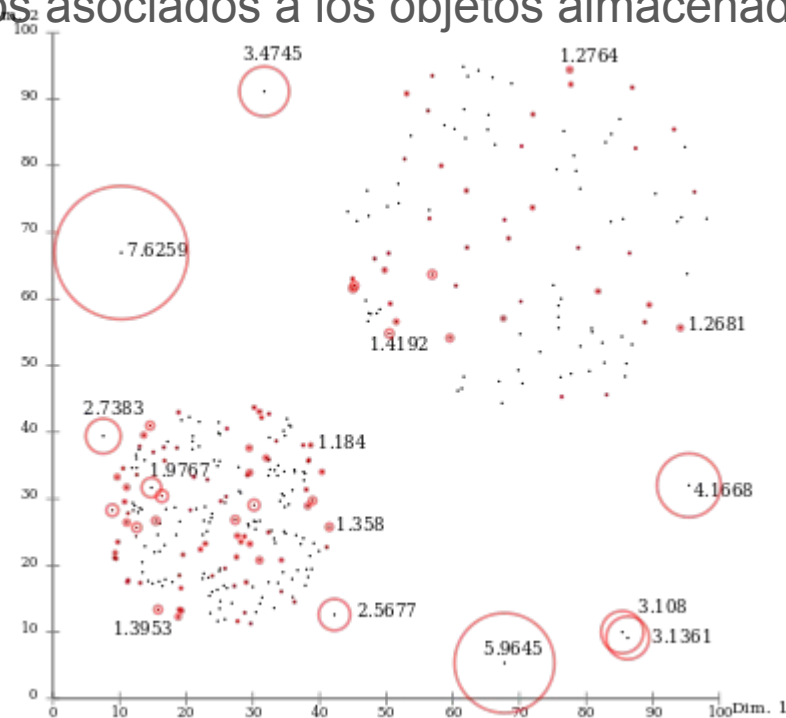


# Outliers

Los **valores atípicos** son datos cuyas características son **considerablemente diferentes** a la mayoría de los otros datos asociados a los objetos almacenados en la base de datos



Son usados para detectar desviaciones significativas a partir de un comportamiento normal (detección de fraudes en tarjetas de crédito)



# Valores perdidos

- La información **no es recolectada** (e.g., error en un sensor de lectura)
- Los atributos **no son aplicables** en ciertos casos

```
titanic_data[titanic_data.Age.isnull()].head(8)
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	244373	13.0000	NaN	S
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	2649	7.2250	NaN	C
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250	NaN	C
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	330959	7.8792	NaN	Q
29	30	0	3	Todoroff, Mr. Lalio	male	NaN	0	0	349216	7.8958	NaN	S
31	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NaN	1	0	PC 17569	146.5208	B78	C
32	33	1	3	Glynn, Miss. Mary Agatha	female	NaN	0	0	335677	7.7500	NaN	Q

# Principales tareas en el pre-procesamiento

Limpieza de datos



Integración de datos



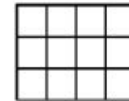
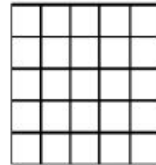
Transformación de datos

-2, 20.00001, 12e-1, 40.0

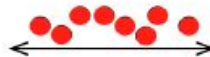


-2.00, 20.00, 1.20, 40.00

Reducción de datos



Discretización de datos



# Limpieza de datos

- Buscar valores **perdidos**
- Identificar **valores atípicos** (outliers) y **ruido**
- Corregir los datos **inconsistentes**
- Resolver la **redundancia** causado por la integración de los datos

# Ejemplo

Encuentre los posibles errores (8) en la siguiente tabla:

ID Cliente	CP	Sexo	Salario	Edad	Estatus	Transacción
1001	32	M	7500	C	C	5000
1002	1C	F	-4000	40	V	4300
1003	31	M	12000000	45	S	2000
1004	0		5400	0	S	7000
1005	27	F	9999	30	D	2100

# Solución

- El cliente 1002 tiene un código postal diferente al resto (contiene letras y generalmente deben ser valores homogéneos)
- El cliente 1004 tiene un código postal de solo 1 cifra (diferente al resto)
- El sexo del cliente 1003 está perdido ;-)
- El cliente 1003 tiene una ganancia de 12e6 por año (parece sospechoso)
- El cliente 1004 tiene una ganancia anual negativa (-40000) y puede verse como un error
- El cliente 1001 tiene un error en su edad (C) posiblemente a causa de una actualización errónea
- 1004 tiene 0 años. Es mejor utilizar fecha de nacimiento en vez de edad (datos calculados)
- Los clientes 1003 y 1004 están Separados o son Solteros? Existe un problema con los valores del atributo "estado marital"

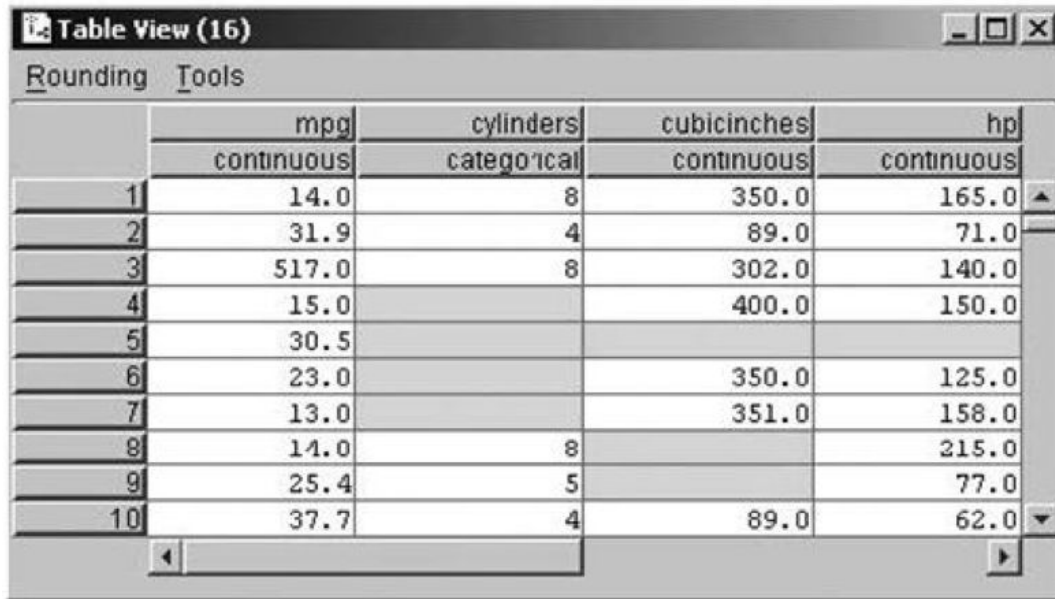
# Qué hacer con los datos perdidos?

- Un método simple es omitirlos sistemáticamente (la o columna)
- No recomendable!, porque puede llevar a un **subconjunto sesgado** de datos y patrones interesantes pueden **perderse**
- Elija un **reemplazo** para los valores de atributos perdidos
  - Utilice valores **constantes** (0, NA, etc.)
  - Utilizar la **media** para valores numéricos y la **moda** para valores discreto (el valor más frecuente)
  - General un **valor aleatorio** (en base a la distribución)



# Ejemplo

La siguiente tabla contiene datos asociados a ciertas características de automóviles que incluyen: gasolina/milla consumida (mpg), el cilindraje (cylinders), centímetros cúbicos (cubicinches) y la potencia (hp).



	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.0	8	350.0	165.0
2	31.9	4	89.0	71.0
3	517.0	8	302.0	140.0
4	15.0		400.0	150.0
5	30.5			
6	23.0		350.0	125.0
7	13.0		351.0	158.0
8	14.0	8		215.0
9	25.4	5		77.0
10	37.7	4	89.0	62.0

Llene los valores faltantes utilizando 1) valores constantes y 2) la media y la moda

# Respuestas

- Caso 1, valores constantes
  - Para valores numéricos, agregar la constante 0:00
  - Para valores categóricos, colocar \NA"
- Caso 2, media y moda
  - Para valores numéricos, agregar los valores 200:65 (cubicinches) y 106:53 (hp)
  - Para valores categóricos, colocar 8
- Caso 3, valores aleatorios

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	8	400.00	150.00
5	30.50	4	144.15	116.55
6	23.00	4	350.00	125.00
7	13.00	6	351.00	158.00
8	14.00	8	323.45	215.00
9	25.40	5	81.84	77.00
10	37.70	4	89.00	62.00

# ¿Qué hacer con los datos perdidos?

- Los tres métodos **no garantizan resultados correctos** (o que los resultados tengan sentido)
- En resumen, reemplazar valores perdidos es un **juego de azar**. Medir los posibles **beneficios contra** la posible **invalidéz** de los resultados

# Transformación de datos

- Las variables tienden a tener intervalos que **varían mucho entre sí**, e.g., la edad, los ingresos mensuales, etc.
- Para algunos algoritmos de minería de datos, las diferencias en los intervalos darán lugar a una tendencia en la que la **variable con mayor intervalo tiene mayor influencia en los resultados**
- Qué hacer?
- **Normalizar** las variables numéricas = estandarizar la escala de cada variable