Ciencia de Datos y herramientas Big Data para la Investigación

TÉCNICAS AVANZADAS PARA EL ANÁLISIS INTELIGENTE DE DATOS

Prof. Dr. Gualberto Asencio Cortés
http://datalab.upo.es/asencio
guaasecor@upo.es

Contenidos

BLOQUE I: Análisis exploratorio de datos
Tema 1: Introducción a la ciencia de datos
Tema 2: Adquisición y visualización de datos
Tema 3: Análisis de la distribución de los datos

BLOQUE II: Preprocesado de datos
Tema 4: Preprocesado de datos
Tema 5: Selección de atributos

BLOQUE III: Aprendizaje no supervisado
Tema 6: Técnicas de clustering
Tema 7: Extracción de reglas de asociación

BLOQUE IV: Aprendizaje supervisado
Tema 8: Técnicas de clasificación
Tema 9: Técnicas de regresión

TEMA 1. INTRODUCCIÓN A LA CIENCIA DE DATOS

Contenidos

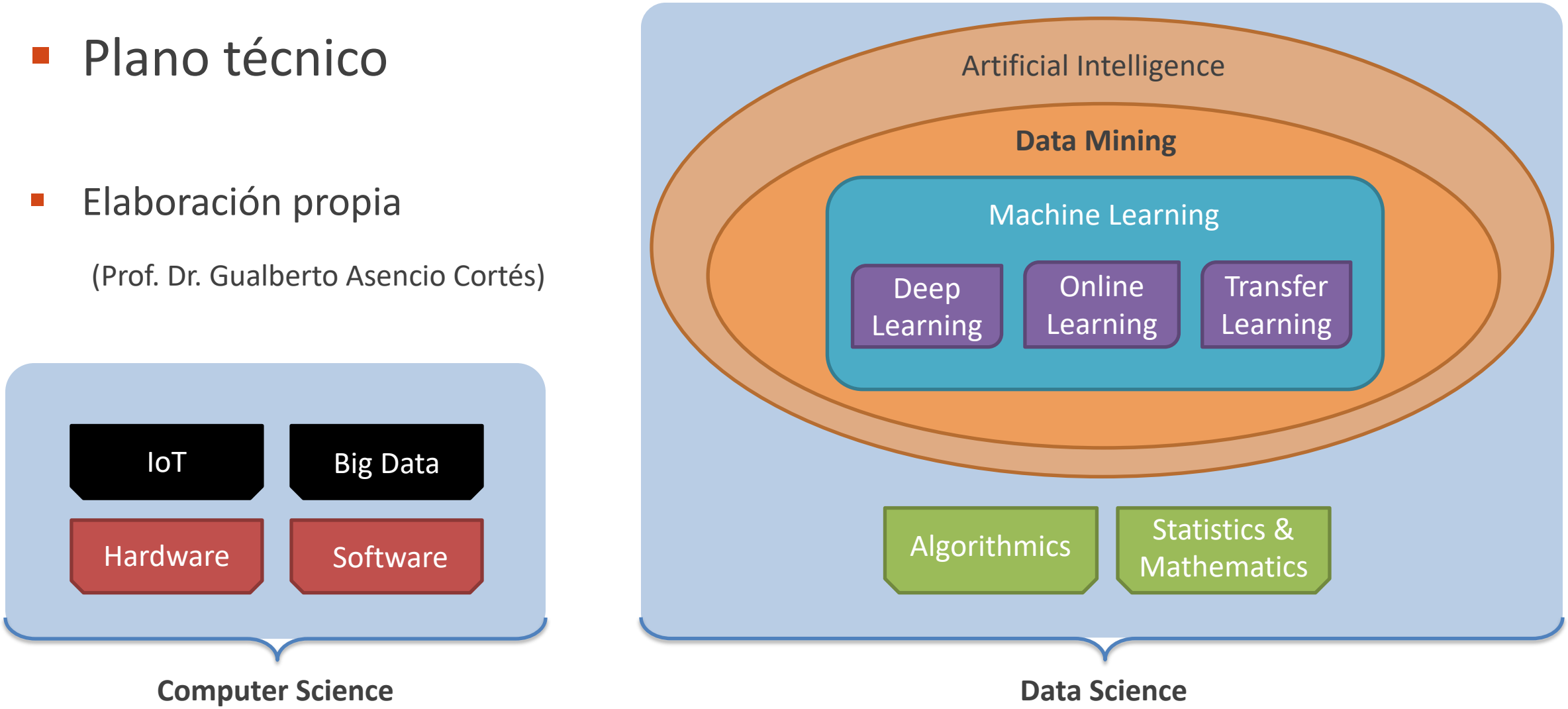
I. Ciencia de datos
II. Conceptos básicos
III. Metodología
IV. Orange

I. Ciencia de datos – Diagrama de conceptos clave

- Plano técnico

- Elaboración propia

  (Prof. Dr. Gualberto Asencio Cortés)

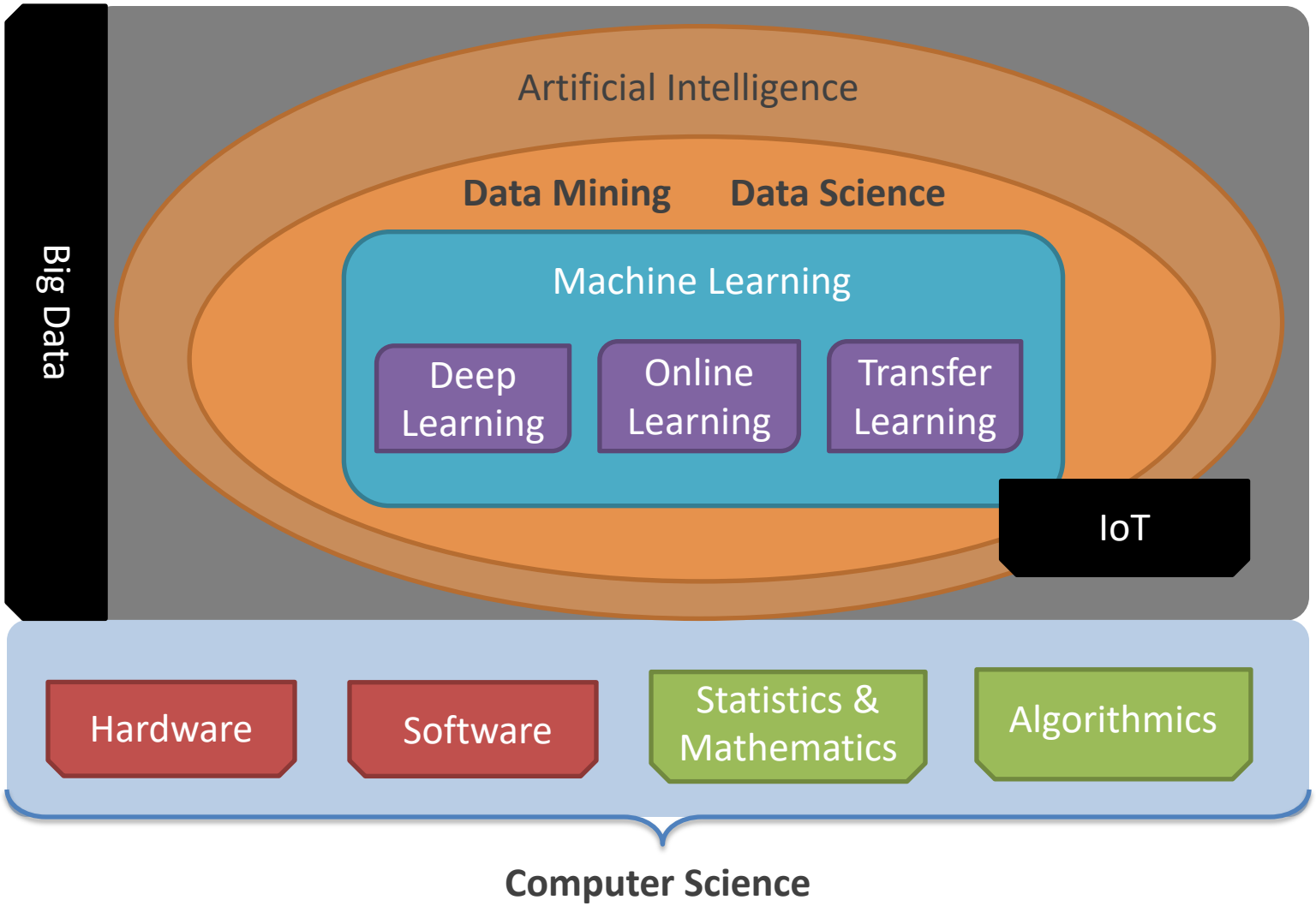


I. Ciencia de datos – Diagrama de conceptos clave

- Plano comercial

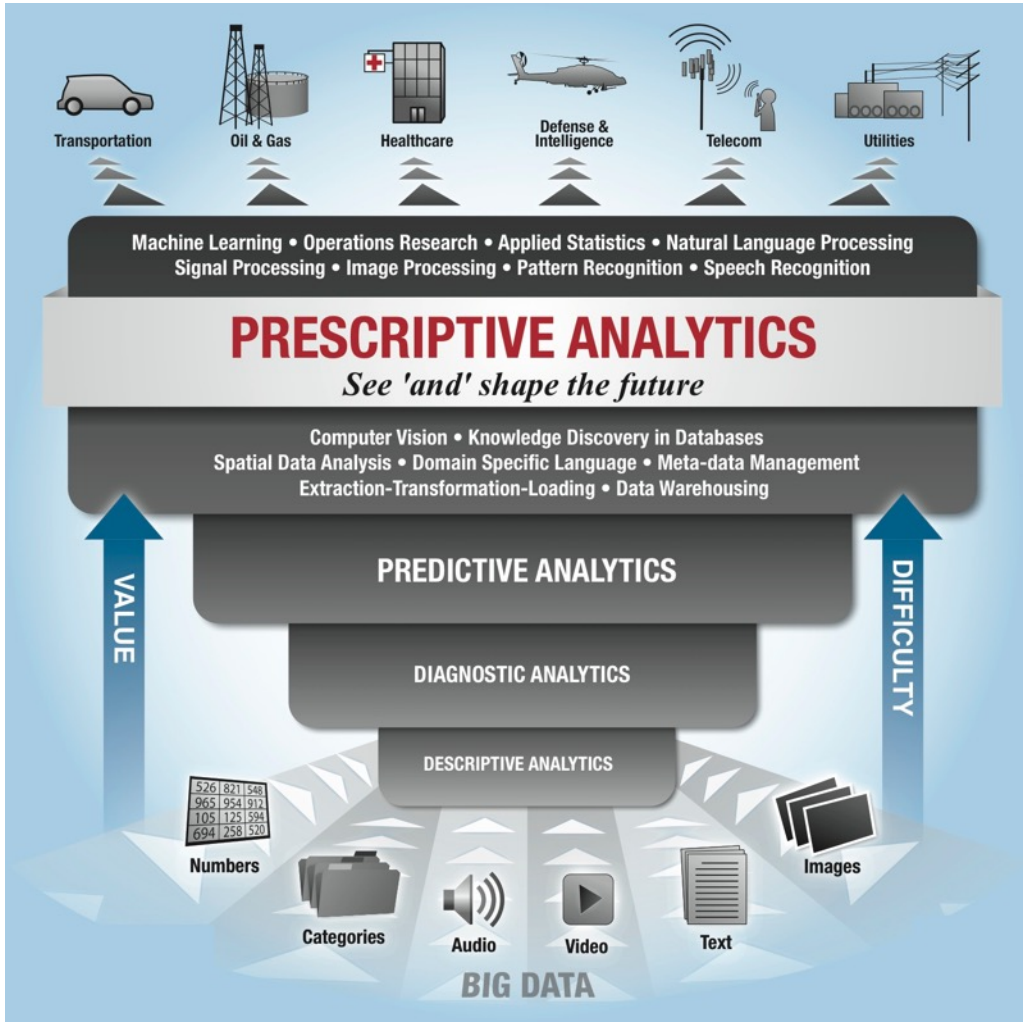
- Elaboración propia

  (Prof. Dr. Gualberto Asencio Cortés)



I. Ciencia de datos – Proyección futura

- Proyección futura del aprendizaje automático



I. Ciencia de datos – Científico de datos

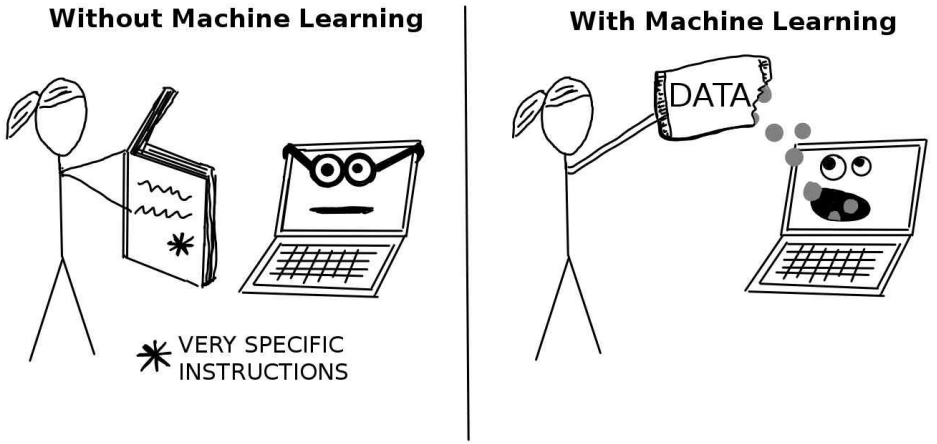
- Es la ciencia que estudia la **extracción de conocimiento** a partir de los datos para obtener información valiosa.
- Roles: el **científico de datos** debe dominar:
  - Tecnologías informáticas
  - Matemáticas y estadística
  - Dominio del problema
  - Capacidad analítica y comunicativa

I. Ciencia de datos – Retos

¿Qué retos podemos abordar?

- Encontrar **perfiles** de clientes fraudulentos (evasión de pagos e impuestos)
- Descubrir **relaciones** implícitas entre síntomas y enfermedades de pacientes de un hospital
- Determinar **relaciones** entre especificaciones técnicas de máquinas, archivos de registro y diagnóstico de errores en fábricas o centros de datos
- **Predecir** el consumo eléctrico en un edificio a varios días-vista
- Estimar la **probabilidad** de que los clientes de una empresa se vayan a la competencia
- Determinar **patrones** de compra de clientes de un supermercado y **recomendar** productos atractivos a clientes en función de lo que compran
- **Segmentar** clientes de forma automática para definir diferentes objetivos de mercado y dirigir campañas de marketing específicas a cada segmento
- … … …

I. ¿Qué es Machine Learning?



- Machine Learning = Aprendizaje Automático
- Programación tradicional:

  Automatizar una tarea usando reglas predefinidas

- Machine Learning: (Idea General)

  Creación automática de reglas

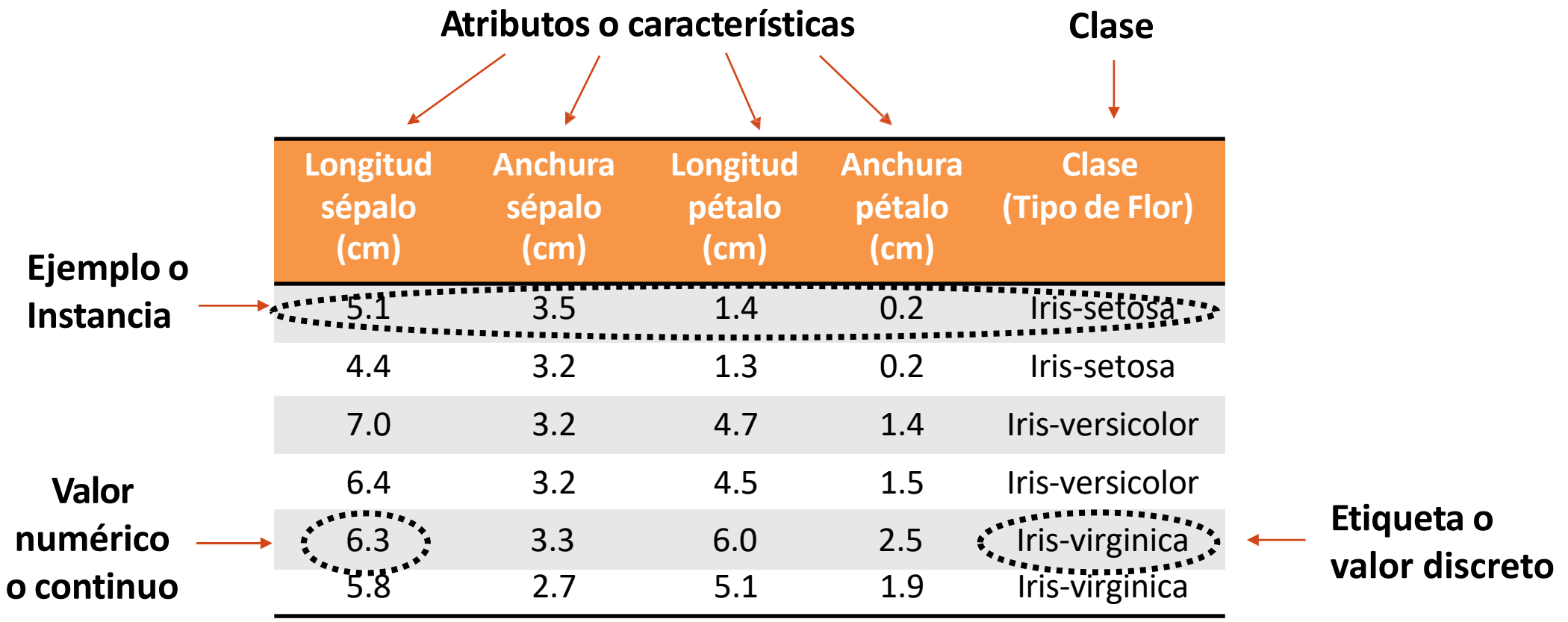
CONJUNTO DE TRAINING

CONJUNTO DE TEST

II. Conceptos básicos

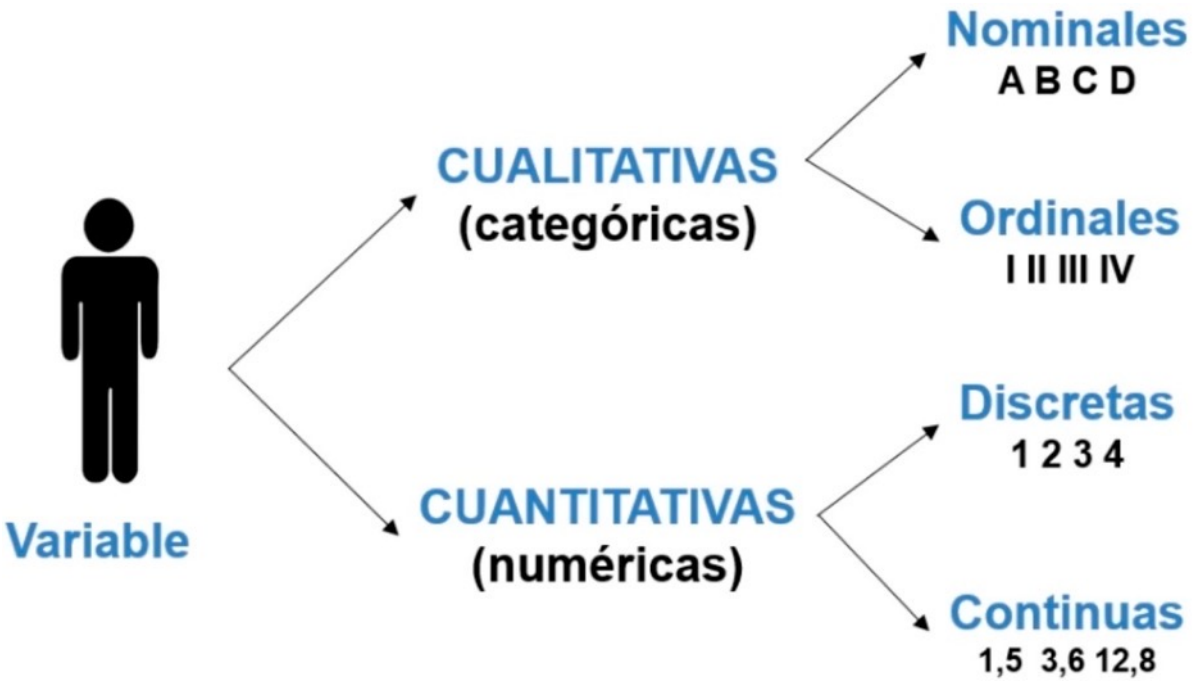
- **Dataset o conjunto de datos:**
  - Filas (ejemplos, instancias, puntos, observaciones, muestras) *(instances)*
  - Columnas (atributos, características) *(features)*
- **Tipos de datos de las columnas:**
  - Valor numérico: valores continuos (cuantitativos)
  - Valor texto o etiquetas: valores discretos (cualitativos)
- **Última columna (opcional):** Clase (atributo de decisión)

II. Conceptos básicos



Longitud sépalo (cm)	Anchura sépalo (cm)	Longitud pétalo (cm)	Anchura pétalo (cm)	Clase (Tipo de Flor)
5.1	3.5	1.4	0.2	Iris-setosa
4.4	3.2	1.3	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica

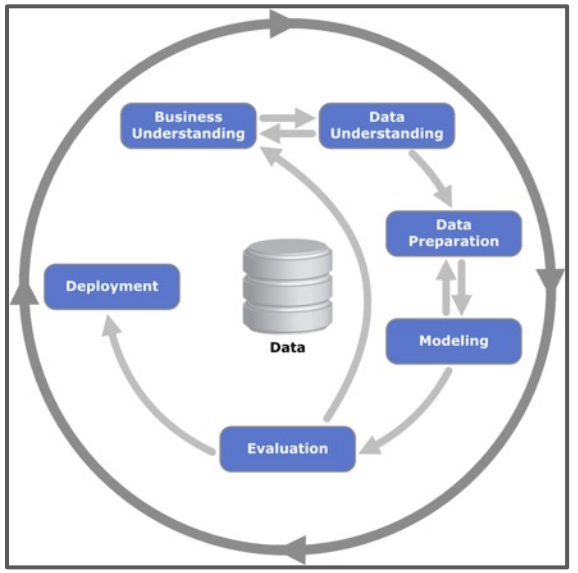
II. Conceptos básicos



III. Metodología

- Metodología para realizar un proyecto de minería de datos.
- **CRISP-DM**:

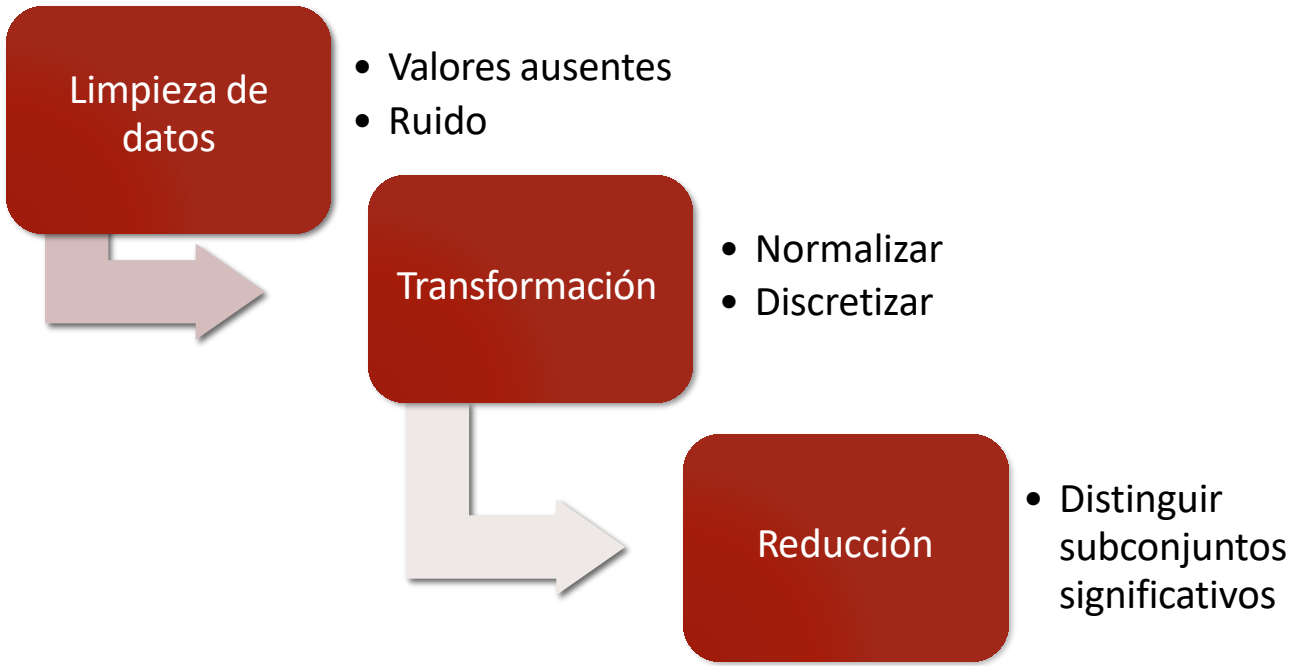
1. Comprensión del negocio
2. Comprensión de los datos
   - Análisis exploratorio de datos
3. Preparación de los datos
   - Extracción, transformación y carga (ETL)
   - Preprocesado de datos
4. Modelado
   - Aprendizaje automático (machine learning)
5. Evaluación de negocio
6. Despliegue



http://crisp-dm.eu

III. Metodología: Preparación de los datos

- Ambigüedades, ruido o no estar en el formato adecuado
- Acelera el algoritmo de aprendizaje
- Mejora la calidad del modelo de conocimiento



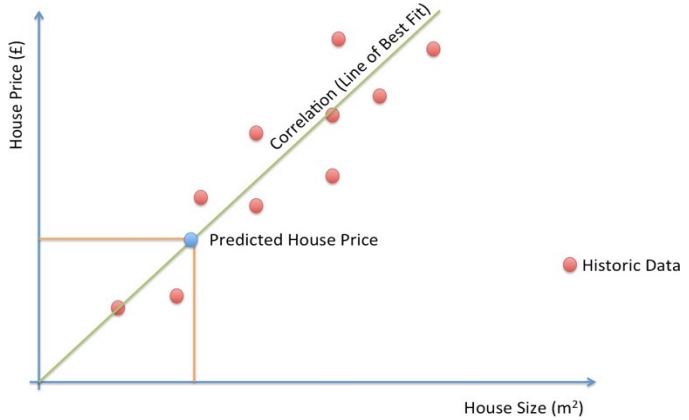
III. Metodología: Aprendizaje

- **Aprendizaje supervisado:**
  - Conjunto de datos previamente clasificado, etiquetado o con un valor numérico asociado.
  - Objetivo: Predecir una clase, etiqueta o valor numérico.
- **Aprendizaje no supervisado:**
  - Conjunto de datos sin clases, etiquetas o valores numéricos asociados.
  - Objetivo: Comprender, resumir, agrupar y encontrar relaciones entre las variables.

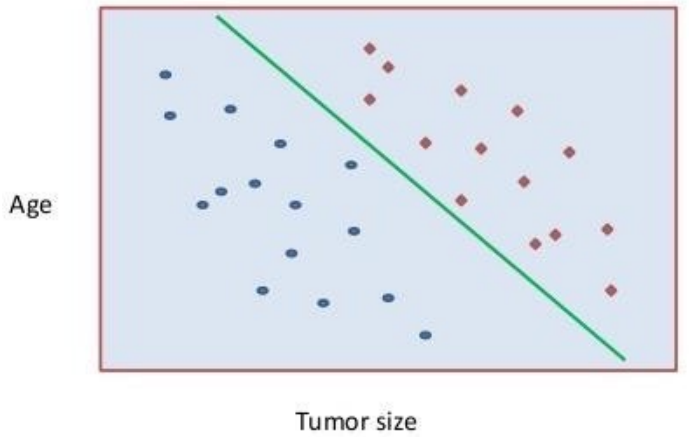
III. Metodología: Aprendizaje

Aprendizaje supervisado

- **Regresión**
  - Predice un dato cuantitativo.
  - Predicción consumo



- **Clasificación**
  - Predice una etiqueta o clase cualitativa.
  - Detección de un determinado tipo de tumor → {benigno, maligno} en función de la edad, tamaño del tumor, densidad, uniformidad de tamaño y forma de célula, etc.



III. Metodología: Aprendizaje

Aprendizaje supervisado

Regresión

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Root.Length
5,0	3,5	1,6	0,6	0,8
4,9	2,4	3,3	1,0	0,3
6,1	2,6	5,6	1,4	0,5
6,1	2,9	4,7	1,4	1,3
4,9	2,5	4,5	1,7	0,6
6,7	3,0	5,0	1,7	1,1
5,9	3,0	5,1	1,8	0,3
5,1	3,3	1,7	0,5	0,5
5,0	3,5	1,6	0,6	0,6
4,9	2,4	3,3	1,0	0,9
...	...	...	...	...
6,2	2,9	4,3	1,3	??
6,3	2,3	4,4	1,3	??
7,7	3,8	6,7	2,2	??
4,8	3,4	1,9	0,2	??

⇨

PREDICCIÓN
0,4
0,3
0,9
0,7

Clasificación

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5,0	3,5	1,6	0,6	setosa
4,9	2,4	3,3	1,0	versicolor
6,1	2,6	5,6	1,4	virginica
6,1	2,9	4,7	1,4	versicolor
4,9	2,5	4,5	1,7	virginica
6,7	3,0	5,0	1,7	versicolor
5,9	3,0	5,1	1,8	virginica
5,1	3,3	1,7	0,5	setosa
5,0	3,5	1,6	0,6	setosa
4,9	2,4	3,3	1,0	versicolor
...	...	...	...	...
6,2	2,9	4,3	1,3	??
6,3	2,3	4,4	1,3	??
7,7	3,8	6,7	2,2	??
4,8	3,4	1,9	0,2	??

⇨

CLASE
versicolor
versicolor
virginica
setosa

III. Metodología: Aprendizaje

Aprendizaje no supervisado

- **Clustering**
  - Agrupación de un conjunto de datos (no etiquetados) en grupos de objetos llamados cluster.
    - Cada cluster está formado por una colección de objetos que son similares (o se consideran similares) entre sí, pero que son distintos respecto a los objetos de otros clusters.

- **Reglas de asociación**
  - Búsqueda de relaciones dentro del conjunto de datos. Establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes;
    - pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros.

III. Metodología: Evaluación

Evaluación:

- Para saber cómo de bueno o malo es el modelo aprendido.
- Medidas:
  - Problemas de **regresión**: diferencia entre valor predicho y valor real.
  - Problemas de **clasificación**: tasa de error.
    - Éxito: La clase de la instancia es predicha correctamente.
    - Error: La clase de la instancia es predicha incorrectamente.
    - Tasa de error: proporción de errores cometidos sobre el conjunto entero de instancias.

III. Metodología: Evaluación

Conjunto de datos para la evaluación:

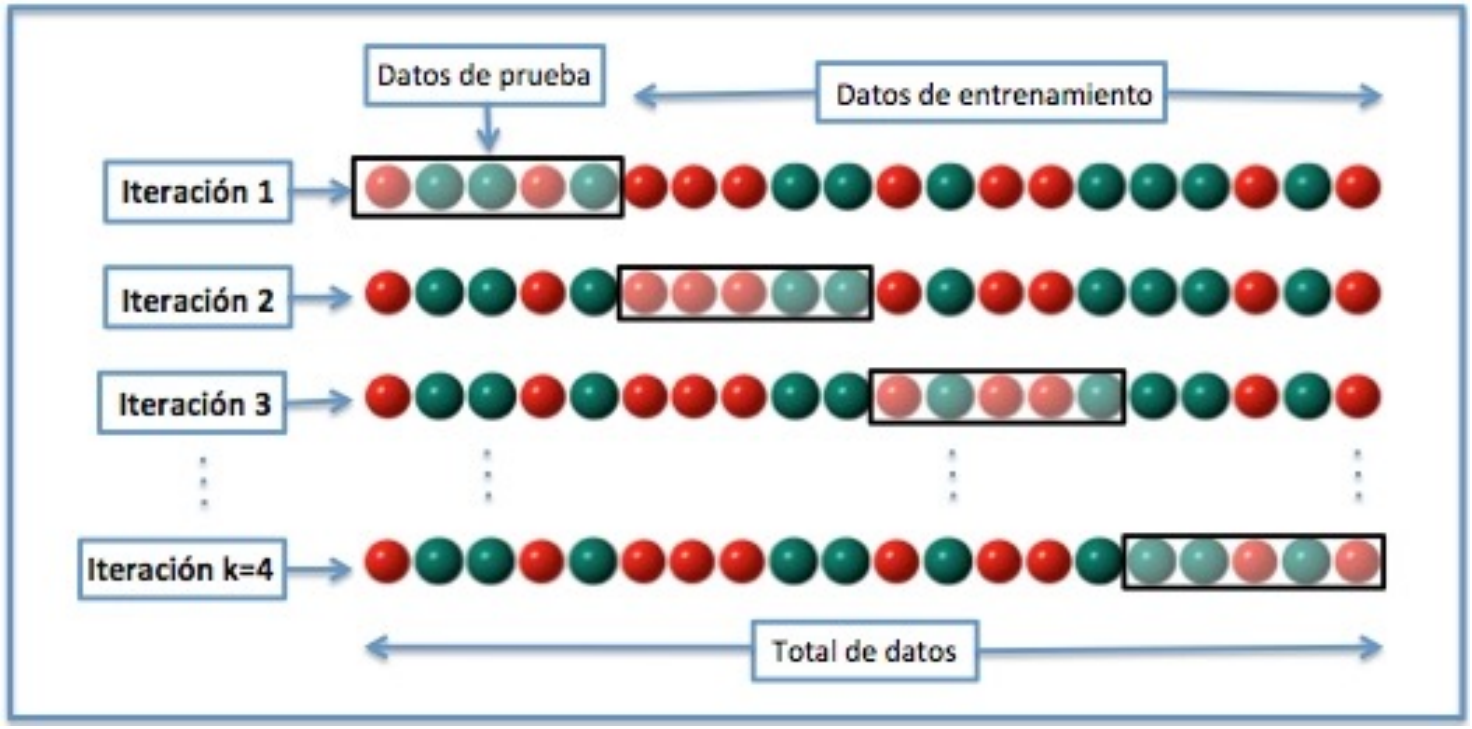
- Debe ser distinto al conjunto de entrenamiento ya que el error que se produzca al clasificar o predecir no es un buen indicador, puede haber sobreajuste del modelo.
- Sets:
  - Training: Entrenamiento de los modelos.
  - Validation: Ajustar los hiperparámetros del modelo con el objetivo de seleccionar el mejor modelo.
  - Test: Evaluación del modelo con registros con los que no se ha entrenado ni ajustado el modelo.

Validación del modelo:

Método hold-out, validación cruzada, leave-one-out, etc.

III. Metodología: Validación

Validación cruzada



Actividad 1

Queremos resolver cada uno de los dos siguientes problemas:

Problema 1: Un supermercado quiere saber el número de productos de un determinado tipo que se venderán la próxima semana.

Problema 2: Un servicio de telefonía quiere saber si un determinado cliente se va a ir a la competencia.

¿clasificación o regresión?

a) Ambos como problemas de clasificación.
b) El problema 1 como un problema de clasificación y problema 2 como un problema de regresión.
c) El problema 1 como un problema de regresión y problema 2 como un problema de clasificación.
d) Ambos como problemas de regresión.

Actividad 1

Queremos resolver cada uno de los dos siguientes problemas:

Problema 1: Un supermercado quiere saber el número de productos de un determinado tipo que se venderán la próxima semana.

Problema 2: Un servicio de telefonía quiere saber si un determinado cliente se va a ir a la competencia.

¿clasificación o regresión?

a) Ambos como problemas de clasificación.
b) El problema 1 como un problema de clasificación y problema 2 como un problema de regresión.
c) El problema 1 como un problema de regresión y problema 2 como un problema de clasificación.
d) Ambos como problemas de regresión.

Actividad 2

Queremos resolver cada uno de los dos siguientes problemas:

Problema 1: Un supermercado quiere saber cuáles son los productos que compran los clientes frecuentemente.

Problema 2: Un servicio de telefonía quiere conocer los grupos de clientes según sus perfiles de consumo.

¿clustering o asociaciones?

a) Ambos como problemas de clustering.
b) El problema 1 como un problema de clustering y el problema 2 como un problema de reglas de asociación.
c) El problema 1 como un problema de reglas de asociación y el problema 2 como un problema de clustering.
d) Ambos como problemas de reglas de asociación.

Actividad 2

Queremos resolver cada uno de los dos siguientes problemas:

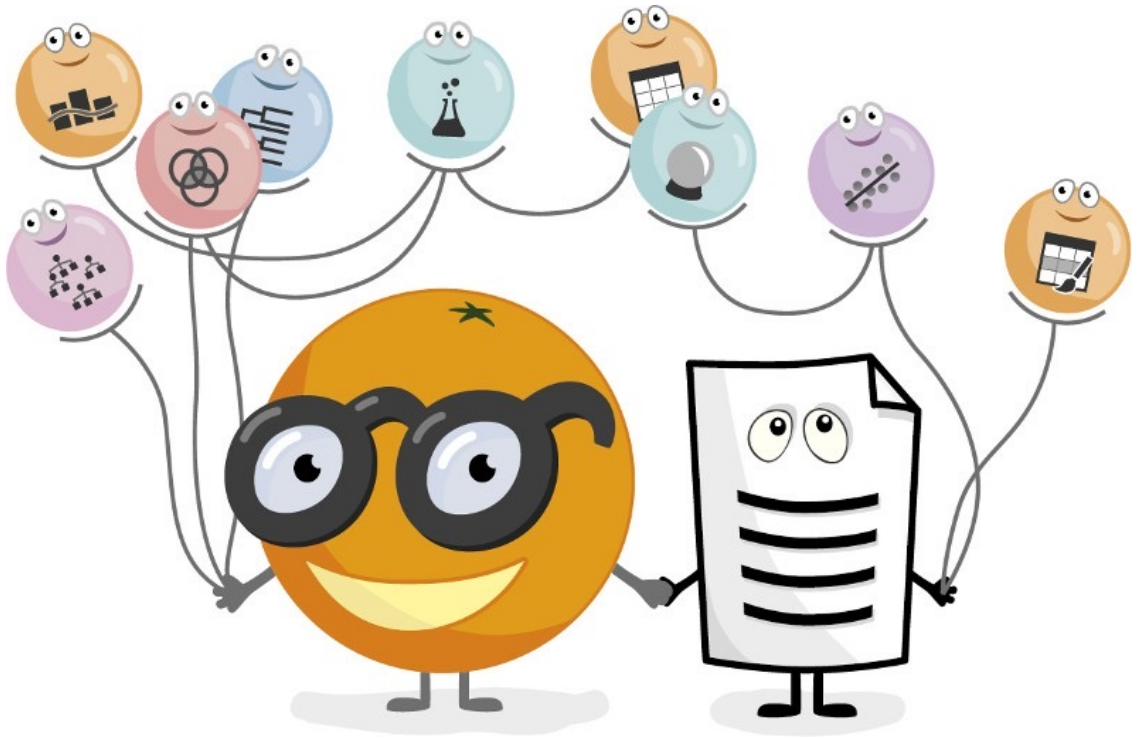
Problema 1: Un supermercado quiere saber cuáles son los productos que compran los clientes frecuentemente.

Problema 2: Un servicio de telefonía quiere conocer los grupos de clientes según sus perfiles de consumo.

¿clustering o asociaciones?

a) Ambos como problemas de clustering.
b) El problema 1 como un problema de clustering y el problema 2 como un problema de reglas de asociación.
c) El problema 1 como un problema de reglas de asociación y el problema 2 como un problema de clustering.
d) Ambos como problemas de reglas de asociación.

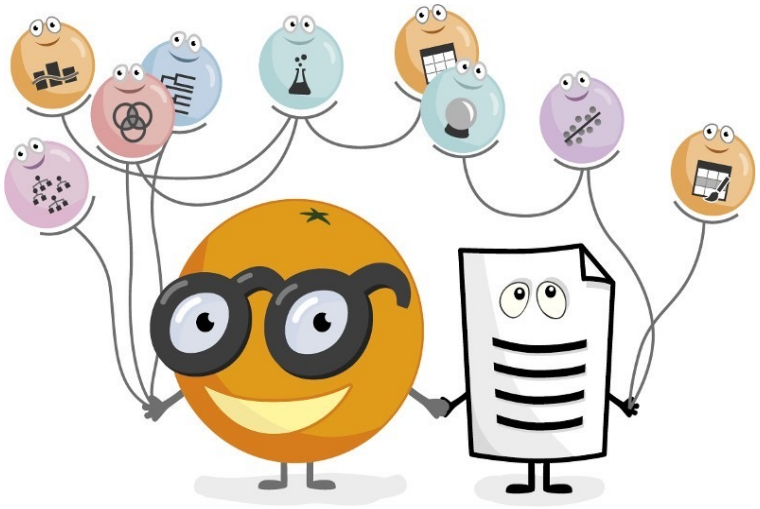
IV. Orange



IV. Orange

Orange es un entorno de trabajo que integra una colección de algoritmos de ciencias de datos.

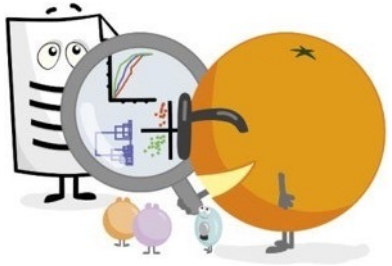
- **Herramientas** para preprocesamiento, clasificación, regresión, clustering, reglas de asociación y visualización.
- Cuenta con un **interfaz visual** muy fácil de usar que evita tener que programar las aplicaciones para nuestros experimentos.
- No se necesita saber **programar.**
- Software open-source y gratuito.
- Descarga: https://orangedatamining.com



IV. Orange

Características:

1) Visualización interactiva

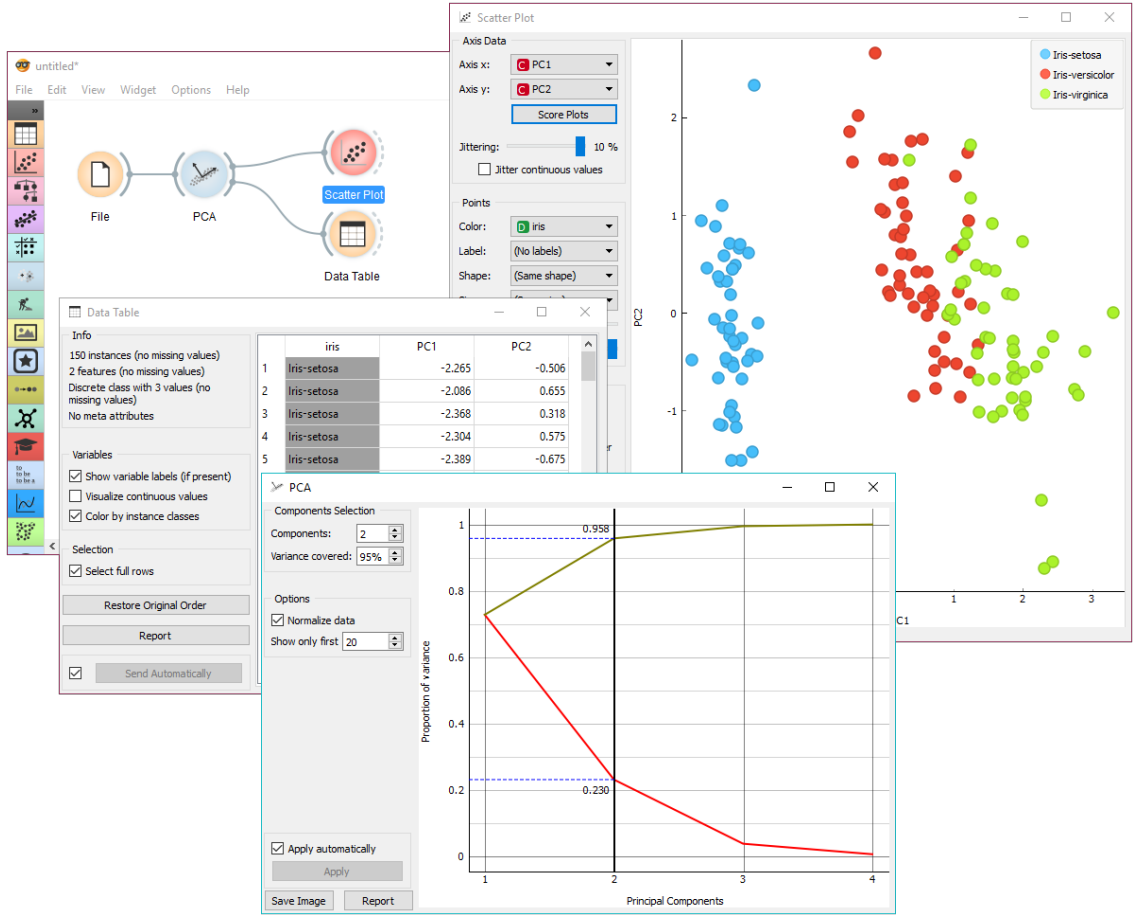


2) Programación visual

3) Add-ons

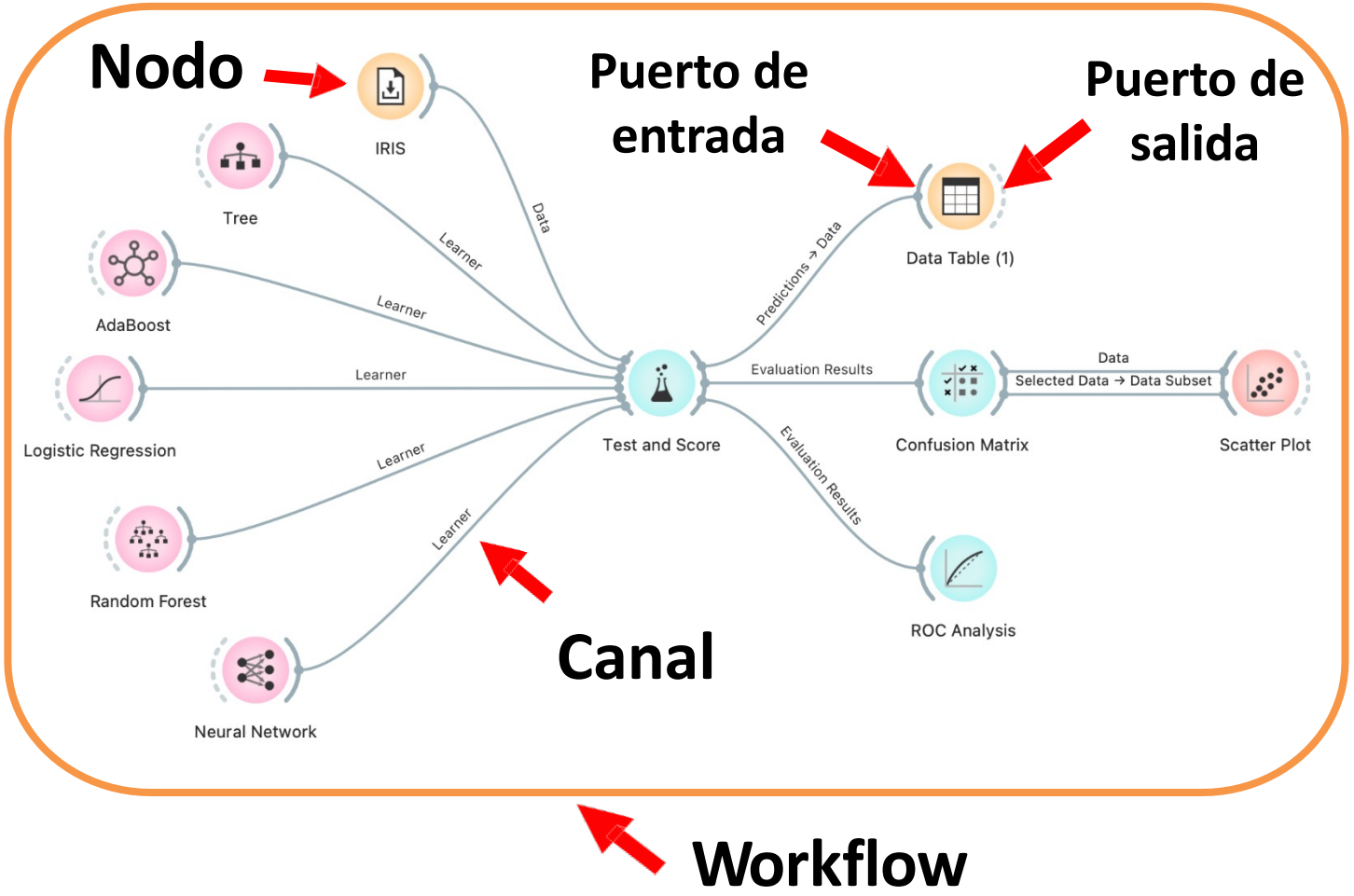
IV. Orange – Visualización interactiva

- Visualizaciones interactivas muy potentes
- Análisis exploratorio de los datos
- Visualizaciones inteligentes
- Informes



Orange – Programación Visual

- Análisis de los datos de forma interactiva
- Minería de Datos basada en **nodos**
- Los nodos se comunican con **canales** a través de **puertos** de entrada y salida.
- Programación basada en flujos de trabajo (**workflows**)



Orange – Add-ons

Instalación de paquetes desde el gestor de paquetes