

Métodos de minería de datos en
Python

Introducción a la minería de datos



Contenido

1

¿Qué es la minería de datos?


2

**Conceptos importantes
en minería de datos**

3

Lenguajes de programación





¿Qué es la minería de datos?

Definición minería

minería

1. f. Arte de laborear las minas.
2. f. Conjunto de los trabajadores que se dedican a la **minería**.
3. f. Conjunto de los facultativos o expertos en **minería**.
4. f. Conjunto de las minas y explotaciones mineras de una nación o comarca.



Definición dato

dato¹ 

Del lat. *datum* 'lo que se da'.

1. m. Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho. *A este problema le faltan datos numéricos.*
2. m. Documento, testimonio, fundamento.
3. m. *Inform.* Información dispuesta de manera adecuada para su tratamiento por una computadora.

Minería de datos

La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados.

Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos estructurando la información obtenida de un modo comprensible para su posterior utilización.

Estos patrones y tendencias se pueden recopilar y definir como
un *modelo de minería de datos*.

https://www.sas.com/es_co/insights/analytics/data-mining.html

<https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions>



UNIVERSIDAD SANTO TOMÁS
PRIMER C. CENTRO UNIVERSITARIO DE COLOMBIA

Minería de datos

Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- Pronóstico: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
- Búsqueda de secuencias: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
- Agrupación: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

Minería de datos

- Riesgo y probabilidad: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
- Recomendaciones: determinación de los productos que se pueden vender juntos y generación de recomendaciones.

Análisis de datos

Análisis de datos descriptivo y exploratorio

- Búsqueda de patrones conocidos
- Mostrar los resultados utilizando técnicas tradicionales

Pros:

- Muchas soluciones
- Más fácil de implementar

Contras:

- No se puede buscar lo inesperado

Minería de datos / Aprendizaje Automático

- Basado en estadística clásica
- El enfoque de la caja negra
- Valores atípicos de salida y correlaciones
- El humano está fuera del circuito
- Preguntas bien definidas que hacer sobre los datos

Pros:

- Escalable

Contras:

- Los analistas tienen que dar sentido a los resultados
- Hace suposiciones sobre los datos

Analítica visual

- Interfaces visuales interactivas
- El humano en el bucle

Pros:

- El ancho de banda visual es enorme
- Identificar patrones desconocidos y errores en los datos

Contras

- La escalabilidad puede ser un problema



¿Qué relación tiene con el big data?

Big Data hace referencia al almacenamiento de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos. Es la tecnología capaz de capturar, gestionar y procesar en un tiempo considerable y de forma eficiente esos datos.

GRANDES DATOS
(BIG DATA)



PANORAMA GENERAL
MUCHAS RELACIONES

MINERÍA DE DATOS
(DATA MINING)



PRIMER PLANO
MUCHOS DETALLES

Big Data el “activo” y Data Mining es el "manejador"



UNIVERSIDAD SANTO TOMÁS
PRIMER C. CENTRO UNIVERSITARIO DE CALI

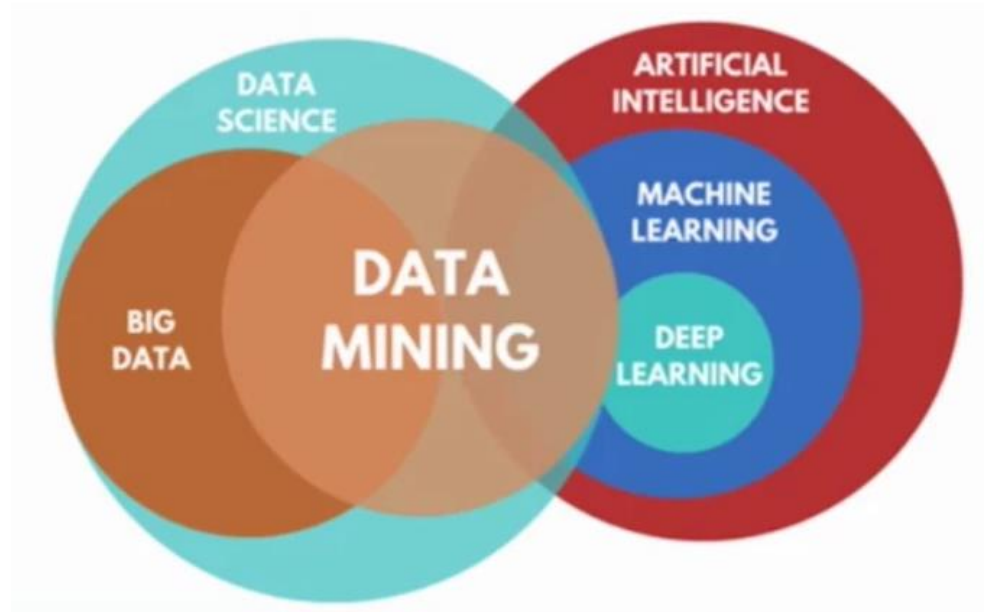
¿Qué relación tiene el aprendizaje de máquina con minería de datos?

Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden **automáticamente**. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros.



ML está orientado hacia el resultado mientras que DM se orienta hacia el descubrimiento de conocimiento.

El maravilloso mundo de la ciencia de datos... e inteligencia artificial.



A horizontal orange bar is positioned to the left of the title text.

Conceptos importantes en minería de datos

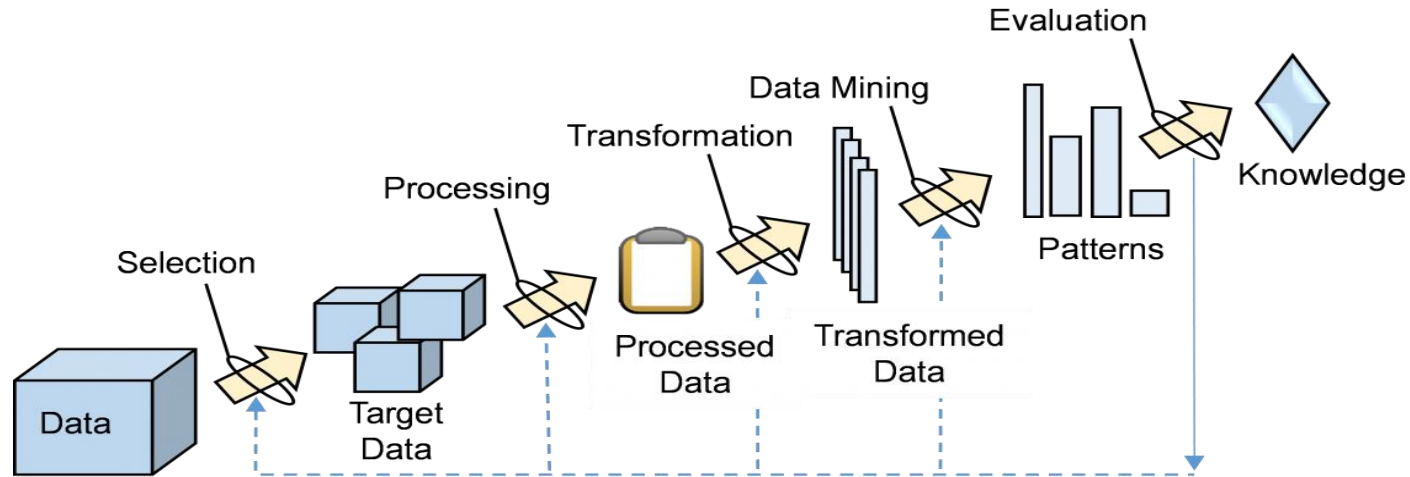
KDD (Knowledge Discovery in Databases)

El descubrimiento de conocimiento en bases de datos se refiere al proceso de identificar patrones válidos, novedosos, potencialmente **útiles** y principalmente **entendibles**.

KDD implica la evaluación e interpretación de patrones y modelos para tomar decisiones con respecto a lo que constituye conocimiento y lo que no lo es. Por lo tanto, el KDD requiere de un amplio y profundo conocimiento sobre tu **área de estudio**.

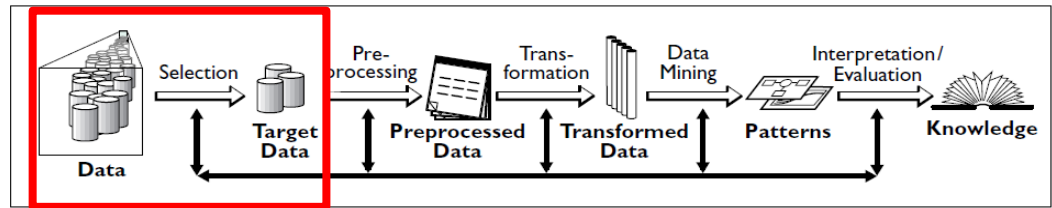


Etapas del proceso KDD



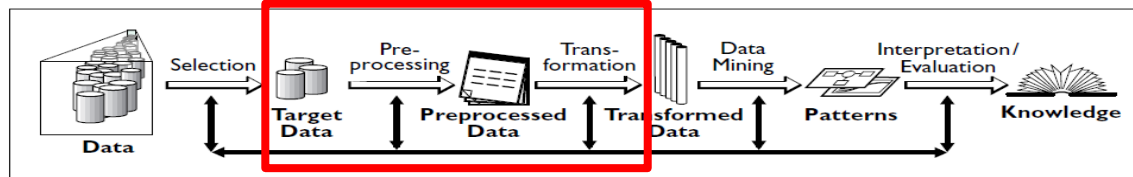
Selección

- Qué tipo de información se requiere?
- Identificación de las fuentes (Relacionales, No Relacionales)
- Cantidad de las fuentes
- Identificación de los tipos de datos (numéricos, texto, etc)



Preprocesamiento/Transformación

- Dar formato a los datos (de bases relacionales a bases funcionales)
- Filtrar los datos (Rangos, valores de interés)
- Editar los datos
 - *Transformar variables*
 - *Crear variables*
- Subconjuntos de datos



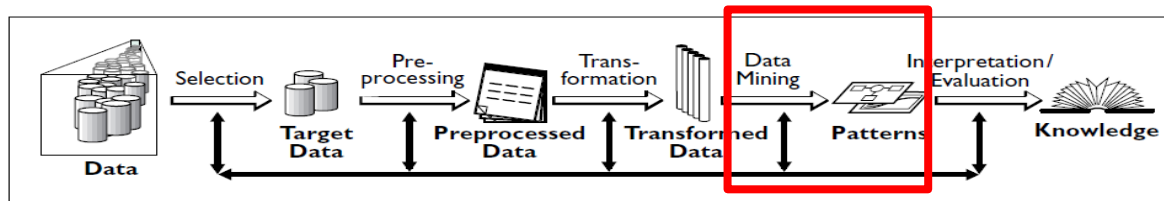
Minería de datos

- Se aplican algoritmos específicos para extraer patrones o relaciones de los datos, los más usuales son:

- Algoritmos de clasificación: Pueden ser interpretables o no

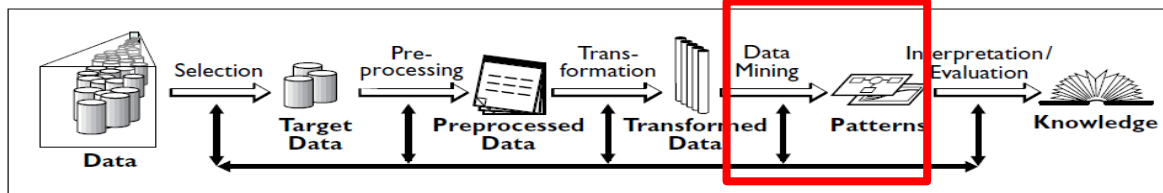
*Arboles de decisión, reglas de decisión

- C4.5, CART, QUEST, RIPPER, CN2 **Interpretables**
- Redes Neuronales, SVM, k-NN, Local Weighted Regression **No interpretables**



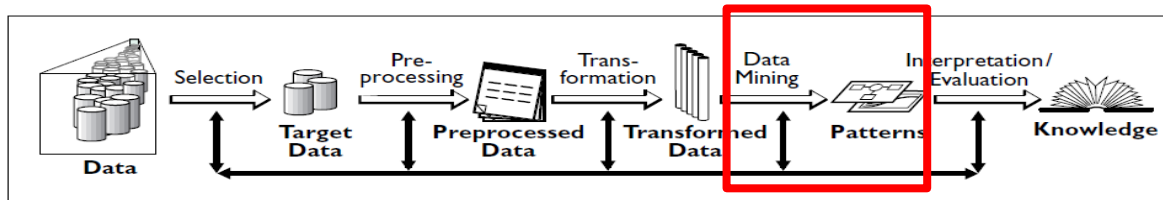
Minería de datos

- Clustering
 - Análisis discriminante (LDA, QDA)
 - K-means
 - Random Forest
 - SVM (Binario)
 - Redes Neuronales
 - GAM, MARS (Binario)



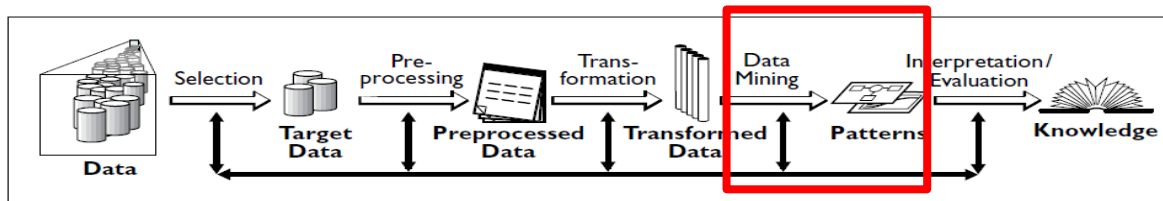
Minería de datos

- Predicción
 - Redes Neuronales
 - Support Vector Regression
 - k-NN
 - PCAR, PLS, Ridge Regression, Lasso Regression, Elastic Net



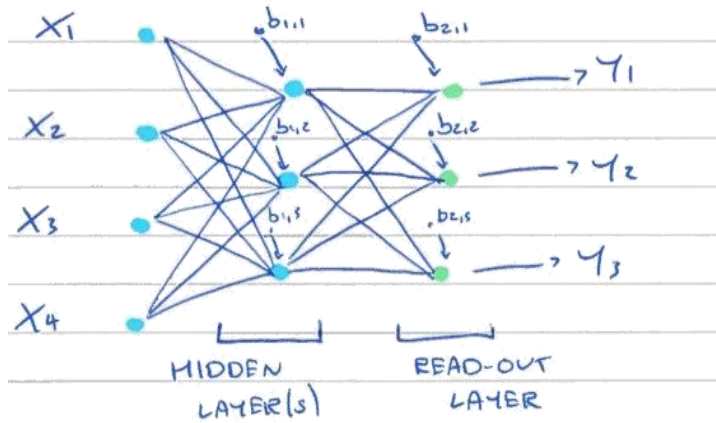
Minería de datos

- Estudio de dependencia
 - Redes Bayesianas
 - Reglas de asociación
 - Cadenas de markov
 - Sistemas de recomendación

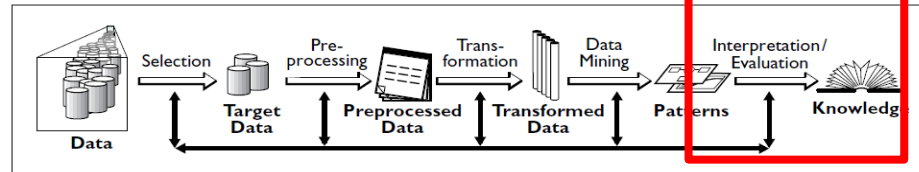
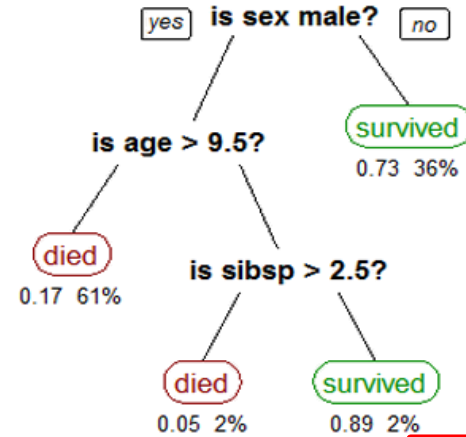


Interpretación/Evaluación

- Sin representación (k-NN, Redes Neuronales)



- Con representación



Aplicaciones

- Negocios / Industria
 - CRM: Segmentación de clientes, caracterización de clientes, evaluación de transacciones, rotación de clientes
 - Detección de fraude
 - Análisis y control de procesos
 - E-commerce, recomendación de productos
 - Análisis del mercado de valores
- Web Mining
 - Text mining, búsqueda y organización documental
 - Análisis de redes sociales



Aplicaciones

- Medicina (reconocimientos de imágenes)
- Farmacología (pruebas de drogas)
- Astronomía (identificación de cuerpos celestes)
- Genética (bioinformática, identificación de patrones de genes)
- Pruebas de satélites (reconocimientos de imágenes)



Casos de éxito

*Una leyenda:
Wal-Mart Cerveza y pañales*



- *los clientes que compran pañales son 5,33 veces más propensos a comprar cerveza (que quienes no los compran)*



Casos de éxito



Data Mining está siendo usada por el Banco Santander para procesar datos de más de 1 millón de clientes, además de trabajar con un gran volumen de informaciones, el data mining consigue actuar con diversas variables.

“Para que tengamos una idea de su capacidad de análisis y proceso, en un trabajo con 15 mil a 20 mil registros, llegamos a operar con 300 variables. Otro ejemplo fue un proceso en el que evaluamos 2 mil variables de cada cliente para medir el comportamiento en nuestra cartera de cheque especial, lo que nos ha permitido comprobar la capacidad de proceso y análisis del producto”, enfatizó el gerente.

Tres meses después del uso de la solución, el Banco Santander ha desarrollado cuatro nuevos modelos de evaluación de riesgo y una serie de otros estudios menores.

https://i1.wp.com/ticsyformacion.com/wp-content/uploads/2013/08/infografia_casos_de_exito_data_mining.png



UNIVERSIDAD SANTO TOMÁS
PRIMER C. CENTRO UNIVERSITARIO DE COLOMBIA

Casos de éxito



https://i1.wp.com/ticsyformacion.com/wp-content/uploads/2013/08/infografia_casos_de_exito_data_mining.png



UNIVERSIDAD SANTO TOMÁS
PRIMER CENTRO UNIVERSITARIO DE COLOMBIA

Lenguajes de programación

¿Cómo hablarle a un computador?

- Instrucciones para fritar un huevo.
- Las instrucciones deben ser:
 - *Claras*
 - *Específicas*
 - *No ambiguas*

¿Cómo hablarle a un computador?

Instrucciones para fritar un huevo:

- Tener a disposición: Una estufa, un fósforo, una paila, una espátula, un huevo, aceite, un plato y sal.
- Prender la llama de la **estufa** con el **fósforo** .
- Tomar la **paila** y colocarla sobre el fogón prendido de la estufa.
- Añadir **aceite** a la paila y dejar que se caliente a 40°C.
- Tomar el **huevo** y golpearlo contra un extremo de la paila hasta que se agriete.

¿Cómo hablarle a un computador?

- Colocar el huevo sobre la paila para terminar de abrir la cáscara y vaciar su contenido dentro de la paila.
- Botar la cáscara en la basura.
- Dejar fritar el huevo por 1 minuto.
- Tomar la **espátula** y sacar el huevo de la paila.
- Colocar el huevo con la espátula en el **plato**.
- Ponerle una pizca de **sal** al huevo.

¿Cómo nos comunicamos los humanos?

Lenguaje:

Conjunto de sonidos y símbolos con un significado

Características:

- **Sintaxis:** Forma del lenguaje.
- **Semántica:** Significado



¿Cómo nos comunicamos los humanos?

Lenguaje:

Conjunto de sonidos y símbolos con un significado

Características:

- Sintaxis: 福
- Semántica: Fortuna

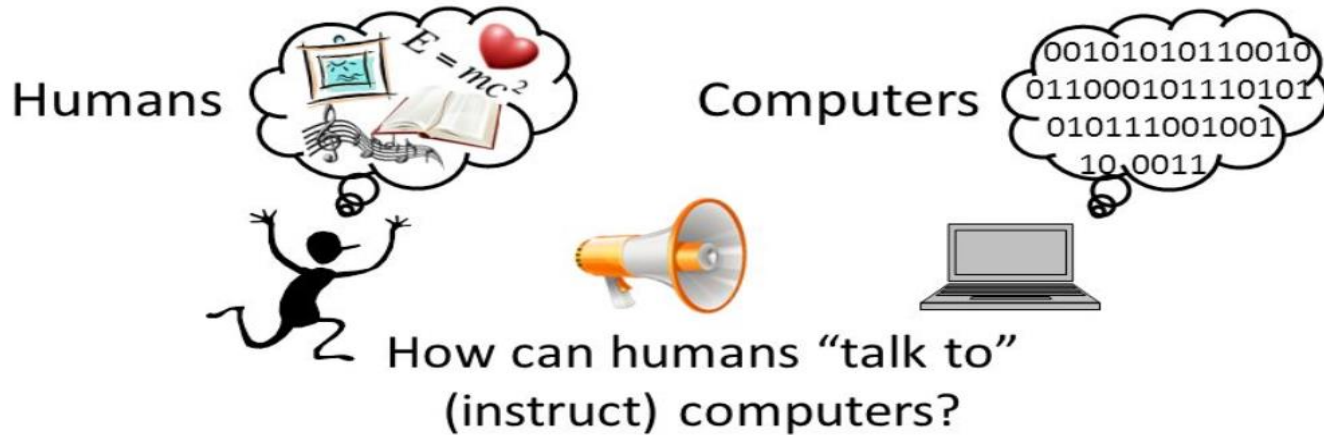


¿Cómo nos comunicamos los humanos?

- Elementos de la comunicación:



¿Cómo hablarle a un computador?



- Impulsos binarios (Sabemos comunicarnos mediante 0's y 1's?)
- Se requirió crear programas para comunicarse con el computador ([ADA LOVELACE](#))

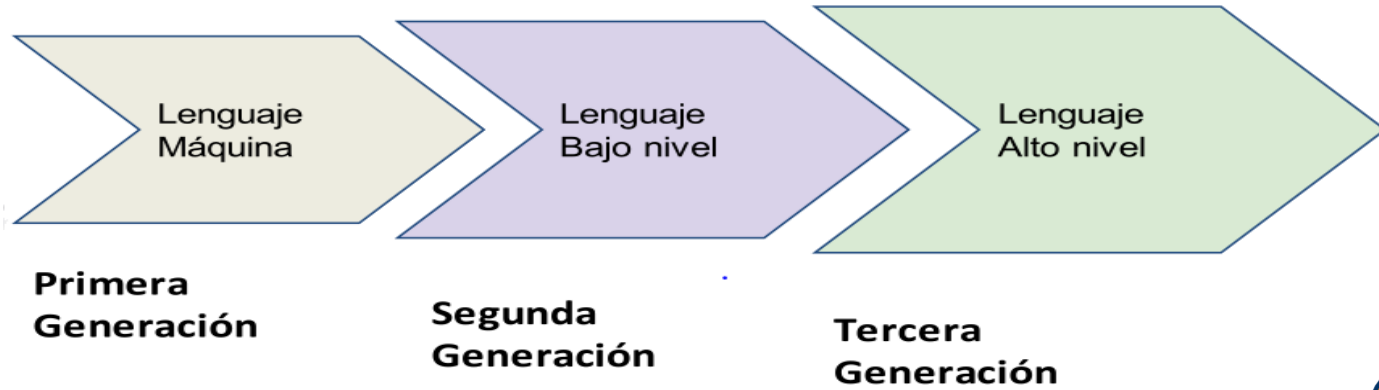
¿Cómo hablarle a un computador?

- A través de programas que permitan comunicarse con él.
 - Los programas deben contener instrucciones claras, precisas y no ambiguas (**Lenguajes de programación**)
- Las instrucciones deben ser:
 - Claras
 - Específicas
 - No ambiguas

Lenguajes de programación

Existe una gran variedad.

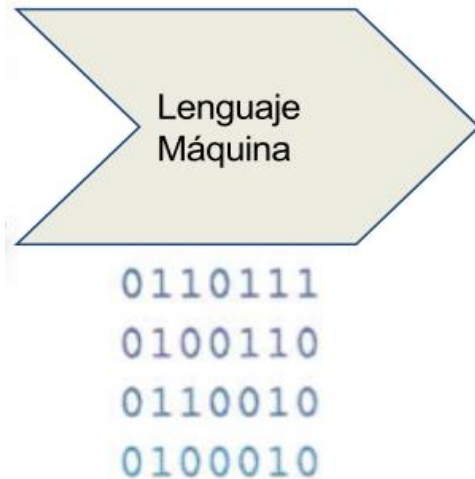
- Clasificación según niveles:



Lenguajes de programación

Existe una gran variedad.

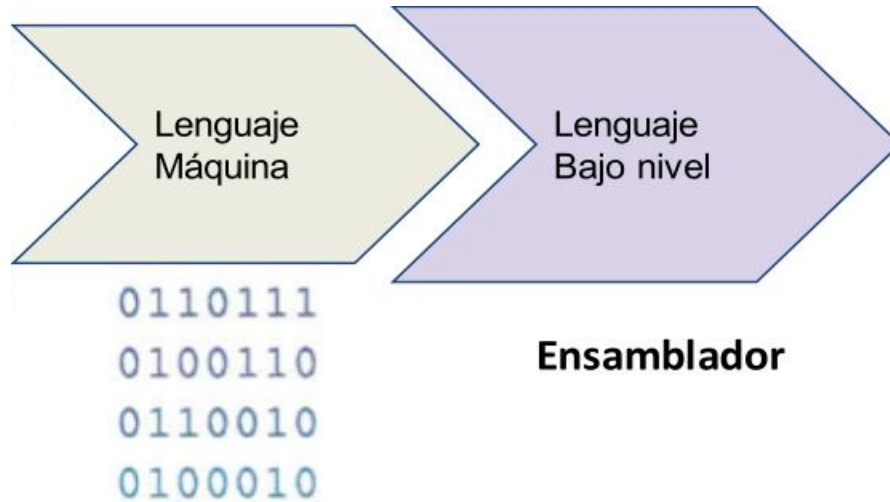
- Clasificación según niveles:



Lenguajes de programación

Existe una gran variedad.

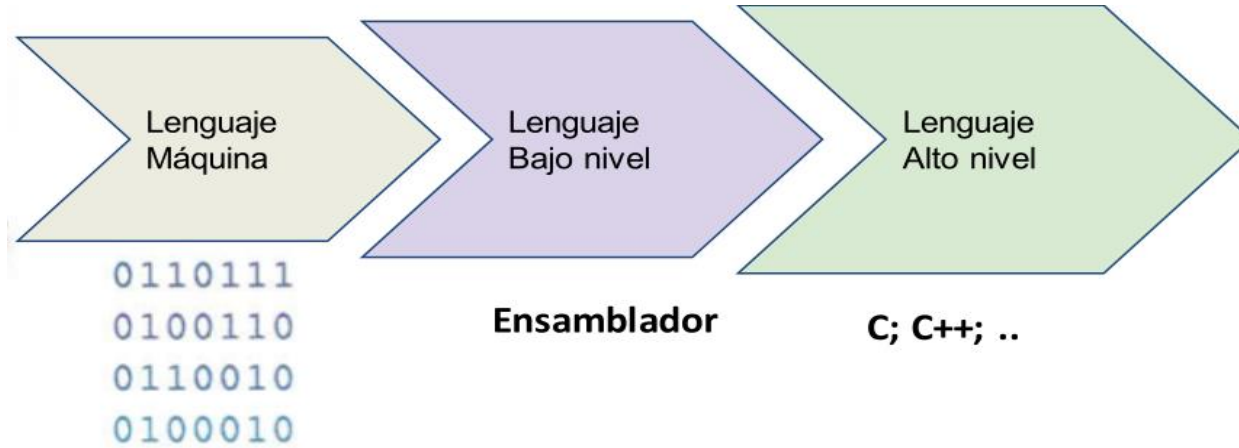
- Clasificación según niveles:



Lenguajes de programación

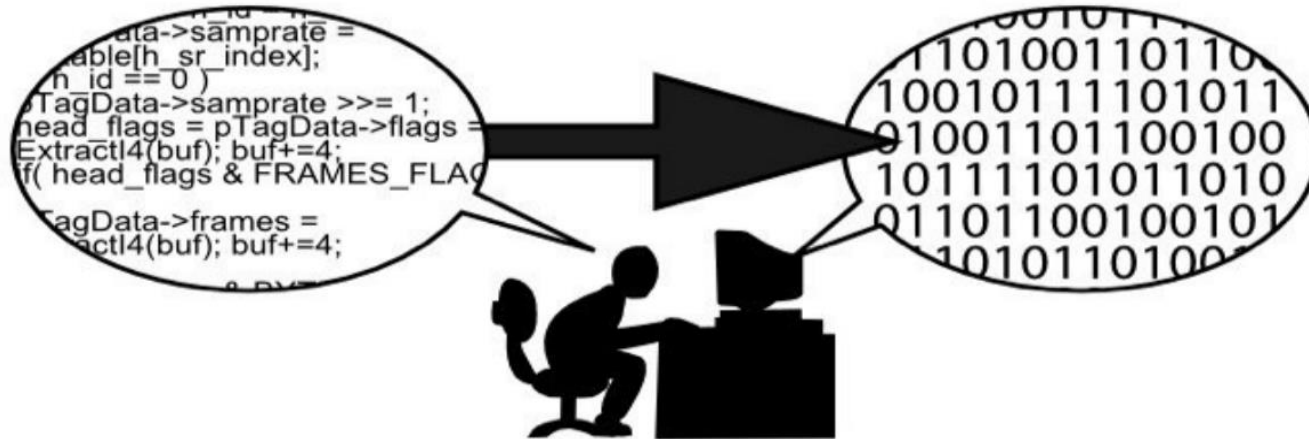
Existe una gran variedad.

- Clasificación según niveles:



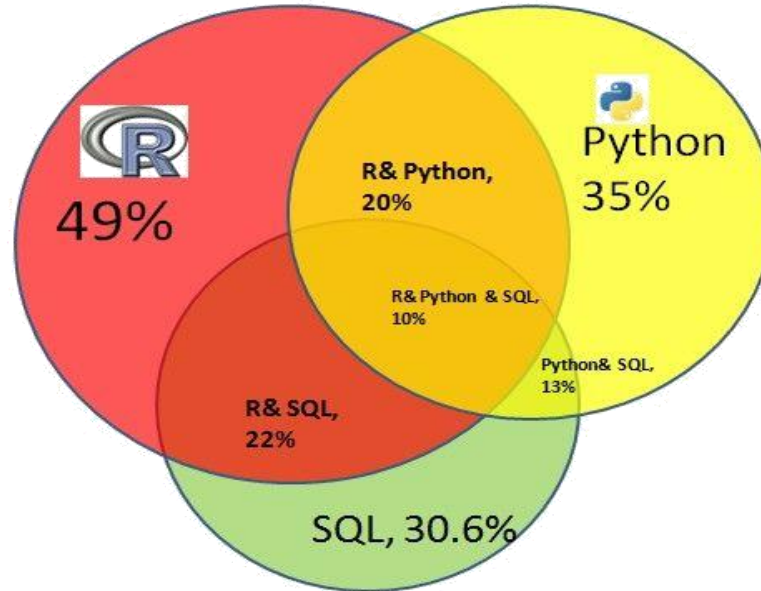
Lenguajes de programación

Sintaxis + Semántica



Lenguajes de programación en ciencia de datos

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining



Gracias

¿Preguntas?

