Introducción a la ciencia de datos y sus aplicaciones

PhD. Carlos Fernando Montoya Cubas

Introducción

Un mundo lleno de datos

Nuestro mundo gira cada vez más en torno a los datos:

- **Ciencia:** astronomía, genómica, medio-ambiente. . .
- **Industria y Energía:** redes de sensores, parques eólicos, previsión de demanda, ciudades inteligentes. . .
- **Ciencias sociales y humanidades:** libros digitalizados, documentos históricos, datos sociales. . .

Un mundo lleno de datos

Nuestro mundo gira cada vez más en torno a los datos:

- **Entretenimiento:** sistemas de recomendación, contenidos digitales, búsquedas multimedia. . .
- **Medicina:** examen de imágenes médicas, previsión de demanda en hospitales, sistemas expertos. . .
- **Financias y negocios:** transacciones de mercados automatizadas. . .

Explosión de datos

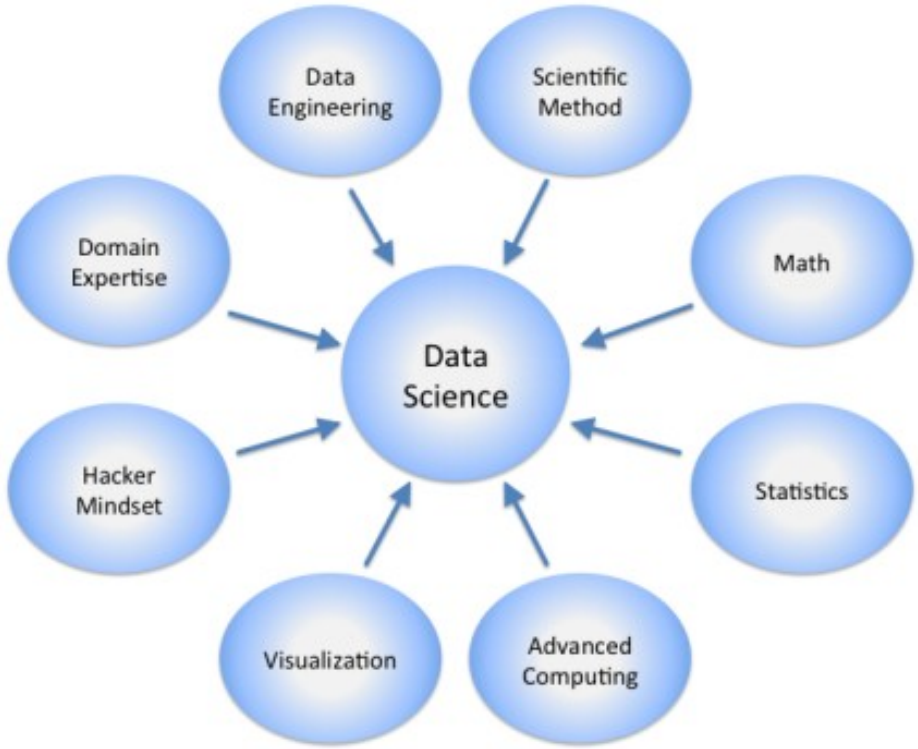
Aunque hace décadas que existen los analistas de datos, también hace décadas que se almacenan datos que no han podido ser procesados hasta hace pocos años:

- Tecnologías de bases de datos
- Coste del hardware de almacenamiento
- Aumento del ancho de banda
- Aumento capacidad de procesado
- Software científico

Definición de ciencia de datos

Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos

Habilidades del científico de datos



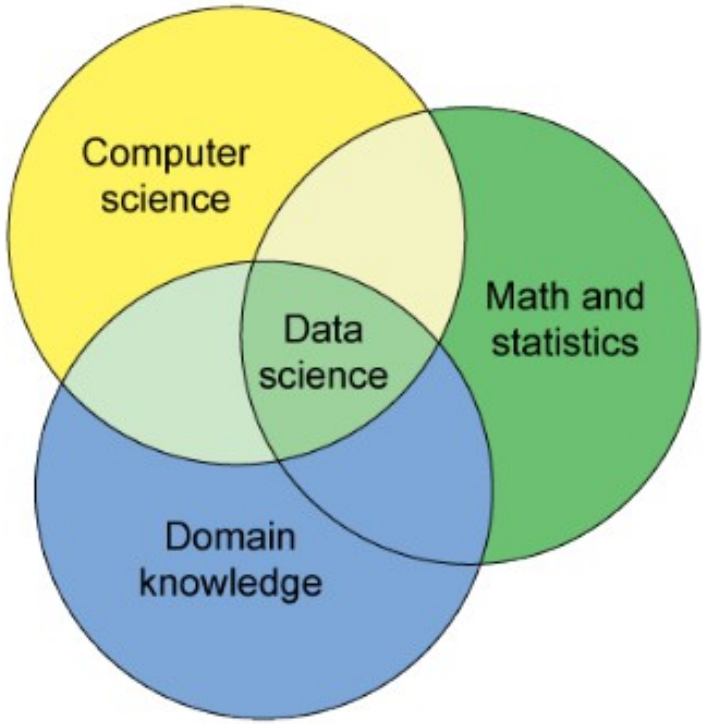


Figura: Fuentes DreamHost e IBM

Minería de datos y KDD

Minería de datos

"La Minería de datos (MD) es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos."

Aunque *Data Science* y *Big Data* son términos más actuales, desde 1989 se denomina a actividades similares como **KDD** (Knowledge Discovery from Databases) o **descubrimiento de conocimiento en bases de datos**.

- El KDD es el **proceso completo de extracción de conocimiento** a partir de bases de datos.

Minería de datos y KDD

- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La Minería de Datos es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

Aportación del término ciencia de datos

Tal vez el término “ciencia de datos” añada más actividades, como por ejemplo el énfasis en la visualización de datos, o el trabajar con datos no estructurados (algo bastante común en el área del big data).

¿Para qué?

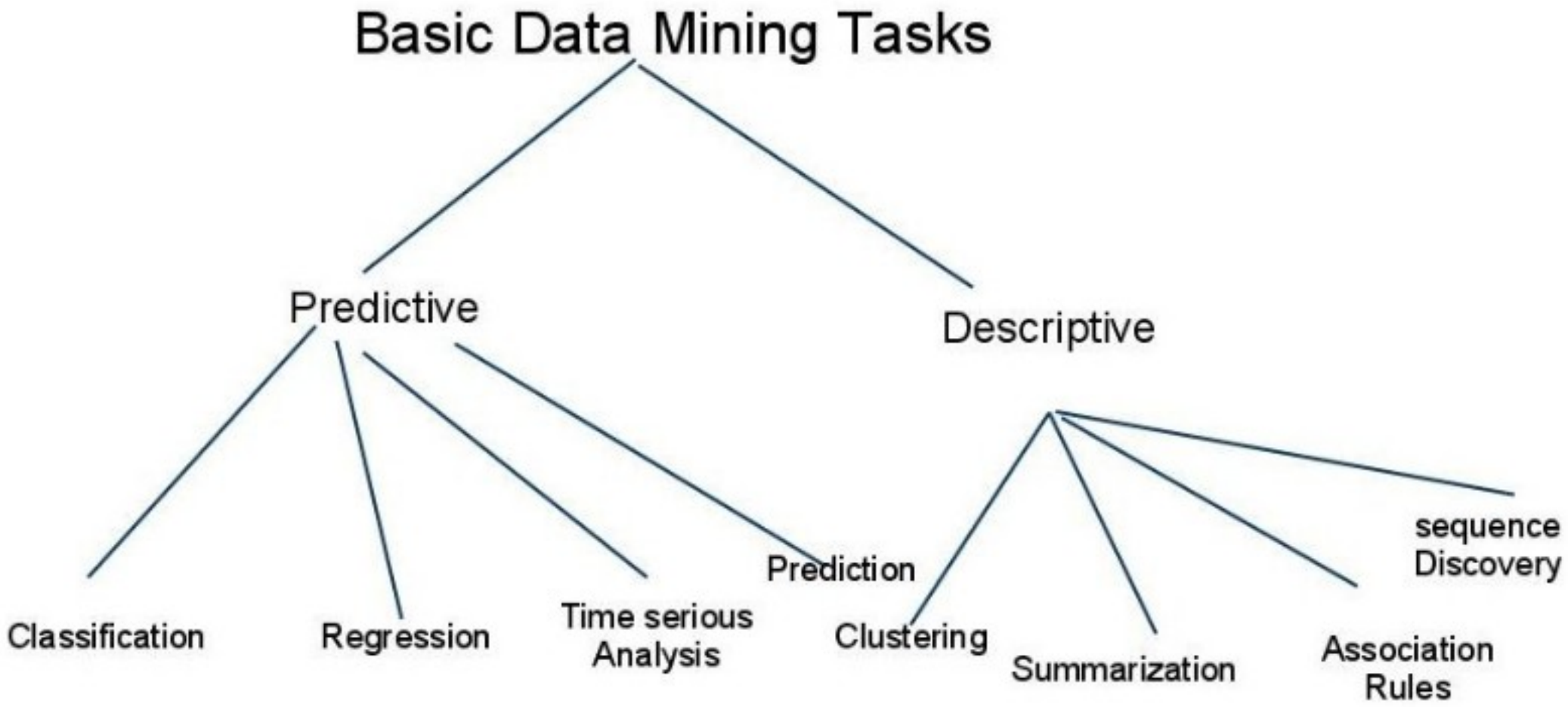
- **Resumir** una gran base de datos
- **Visualizar** datos multi-dimensionales
- **Predecir** valores ⇒ Nos centraremos en este.
- **Explicar** los datos existentes

Orígenes de datos

Las fuentes de datos son muy variadas, a menudo incluso se mezclan, dando lugar a disciplinas como fusión de información, extracción de características, preprocesamiento de datos:

- Bases de datos relacionales
- Bases de datos espaciales y/o temporales: satélites, redes de sensores, telefonía móvil (Cómo te espía tu centro comercial por WiFi y BlueTooth)
- Bases de datos de documentos: archivos históricos...
- Bases de datos multimedia: imágenes, vídeos, sonidos. . .
- La World Wide Web
- Grandes volúmenes de datos no estructurados (Big Data)

Técnicas de Minería de Datos



¿Cómo extraer conocimiento?

La Ciencia de Datos trata de extraer conocimiento de los datos mediante:

- Técnicas estadísticas clásicas
- Inteligencia Artificial y Aprendizaje automático

Muchos métodos de aprendizaje automático se apoyan en **métodos de optimización** matemática y **técnicas estadísticas**, sin embargo a menudo se combinan con técnicas de inteligencia artificial para superar las limitaciones de los primeros en cuanto al entrenamiento de modelos, pero también para diseñar soluciones a problemas, crear sistemas expertos, etc.

¿Por qué IA?

En este nuevo milenio:

- La ciencia y la tecnología están cambiando rápido.
- Se tiene relativamente bastante conocimiento de distintos campos de la ciencia más tradicionales (p. ej. física).
- Los computadores están extendidos por todo el mundo.

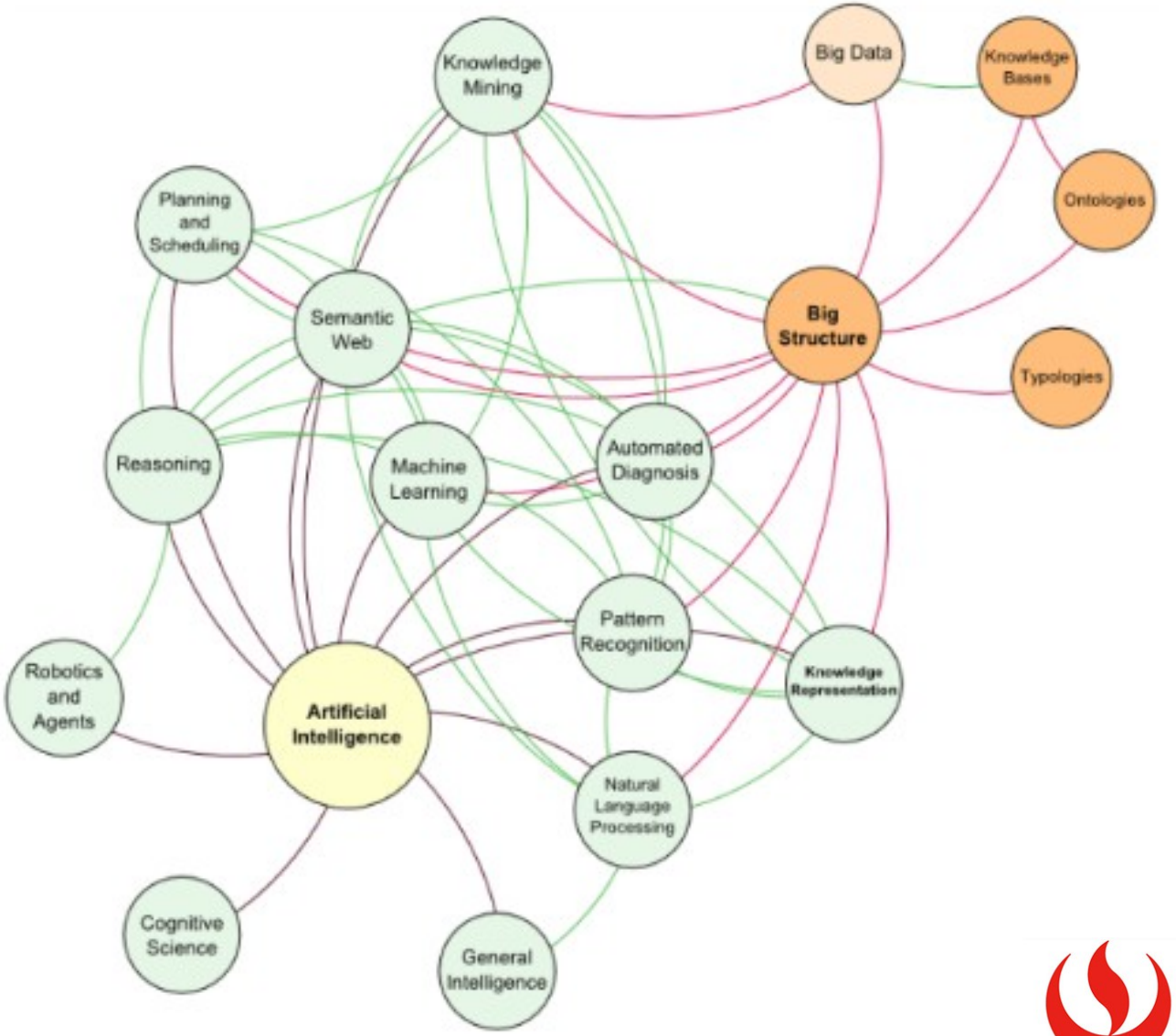
Grandes retos de la ciencia y la tecnología:

- **Comprender el cerebro** (razonamiento, conocimiento, creatividad).
- **Crear máquinas inteligentes:** ¿Es esto posible? ¿Cuáles son los retos tecnológicos y filosóficos?
- IA presenta las preguntas y retos más interesantes de la informática en la actualidad.

¿Qué se necesita?

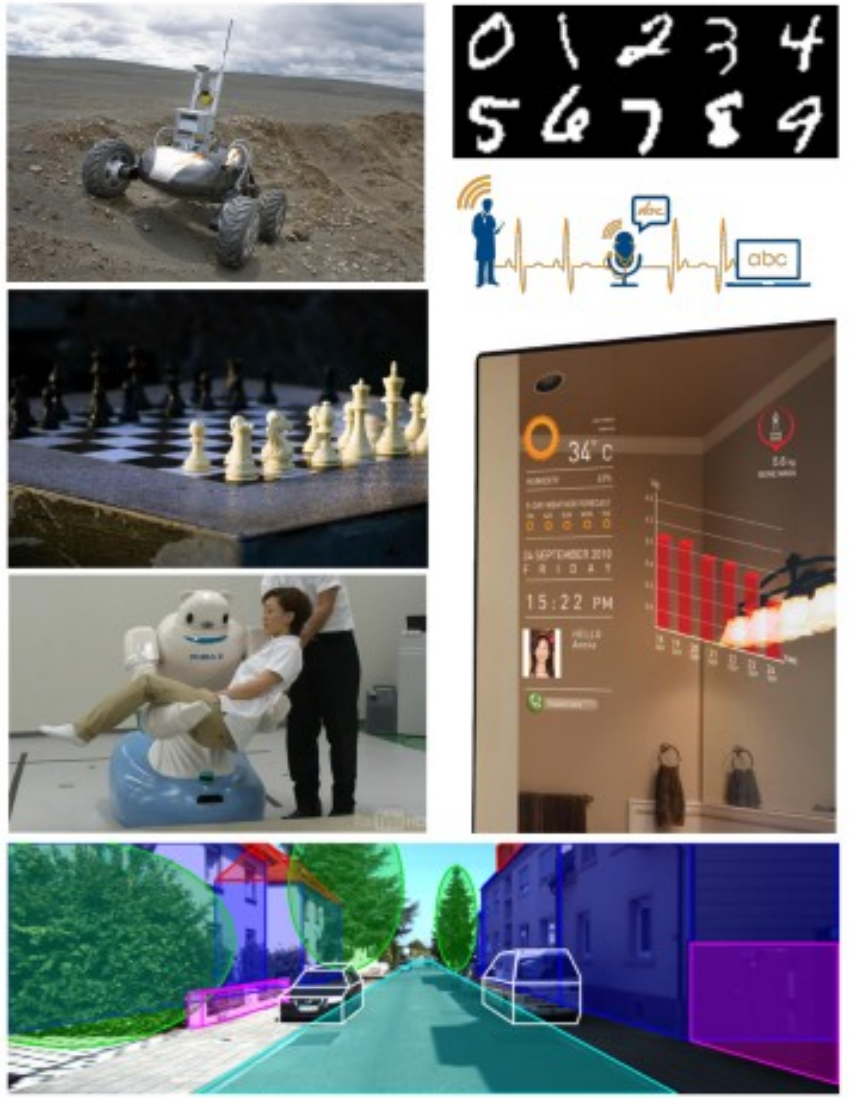
Conducción autónoma
Visión por computador, detección de obstáculos, análisis
de señales de tráfico, mecanismo
de control del vehículo, planificación de rutas, evaluación
del terreno, etc.

Proyecto Proverb
Procesamiento del lenguaje natural, conocimiento extenso
del lenguaje, la historia y la cultura popular, búsqueda de soluciones posibles.



Aplicaciones de la IA

- Navegación autónoma
- Tecnologías asistidas
- Detección de objetos
- Reconocimiento de escritura/habla
- Planificación estratégica
- Inteligencia ambiental
- Sistemas de recomendación
- Medicina
- Diseño industrial



UPC
Universidad Peruana de ciencias Aplicadas

¿Cuándo usar IA?

Son tareas de gran impacto social, diversas y complejas.

- No exista una solución analítica o algorítmica conocida.
- Cuando existan demasiadas posibilidades que hagan difícil el cómputo y podamos usar heurísticas para reducirlo.
- Cuando es difícil el tratamiento de la información y posiblemente sea incompleta o imprecisa.

Aprendizaje automático

Machine learning

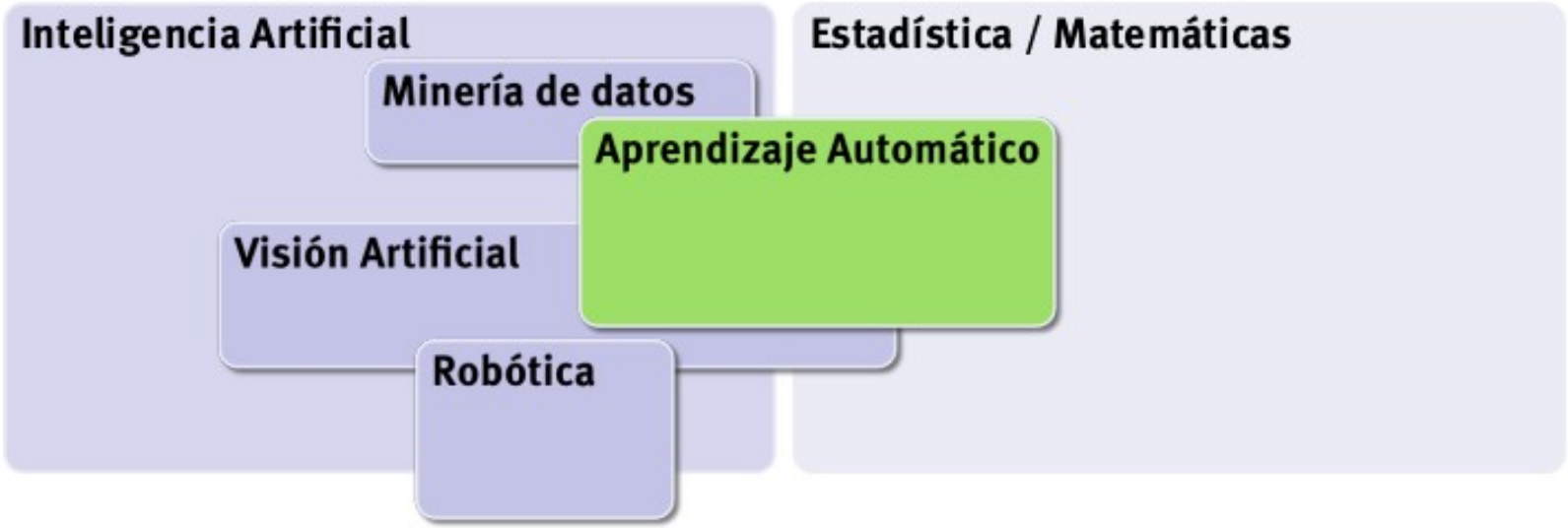
El aprendizaje automático o aprendizaje máquina (*machine learning* en inglés) se define como “campo de estudio que proporciona a los ordenadores la capacidad de aprender sin haber sido explícitamente programados”.

El aprendizaje automático equivale a “aprender de los datos” con el fin de extraer el conocimiento necesario según diferentes propósitos.

Este **“aprender de los datos”** hace que el aprendizaje automático se sitúe entre diferentes ramas que pertenecen a la inteligencia artificial, la estadística y las matemáticas

Aprendizaje automático

Área de estudio que confiere a los ordenadores (máquinas) la habilidad de aprender sin haber sido específicamente programadas para la tarea en c



Aprendizaje automático

- El aprendizaje automático como herramienta para examinar grandes repositorios de datos de Big Data.

Objetivo

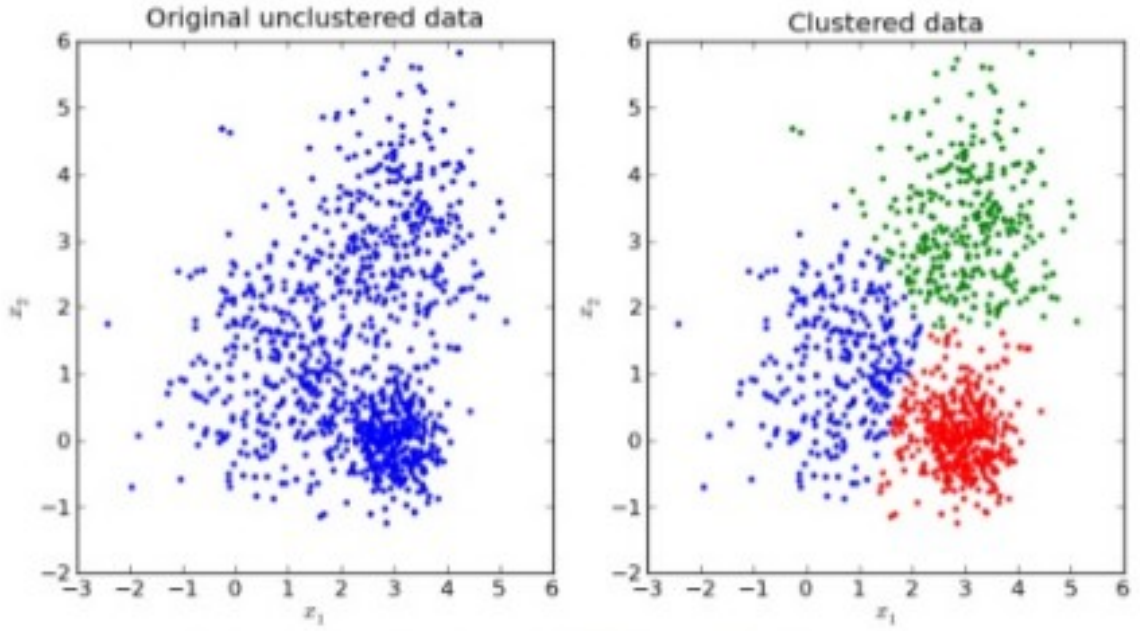
Ayudar en la toma de decisiones descubriendo patrones ocultos, relaciones desconocidas, predicciones y otra información útil ⇒ ventajas competitivas para las empresas que lo posean.

- “Algunos analistas confirman que las empresas que adopten técnicas de analítica de Big Data tendrán una ventaja competitiva de 20 % en todas las métricas financieras sobre sus competidores” Gustavo Tamaki (2012 “La hora del Big Data”).

Descripción: agrupamiento

K-means Clustering

- partition ***n*** observations into ***k*** clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- http://en.wikipedia.org/wiki/K-means_clustering



http://pypr.sourceforge.net/kmeans.html

Predicción: regresión

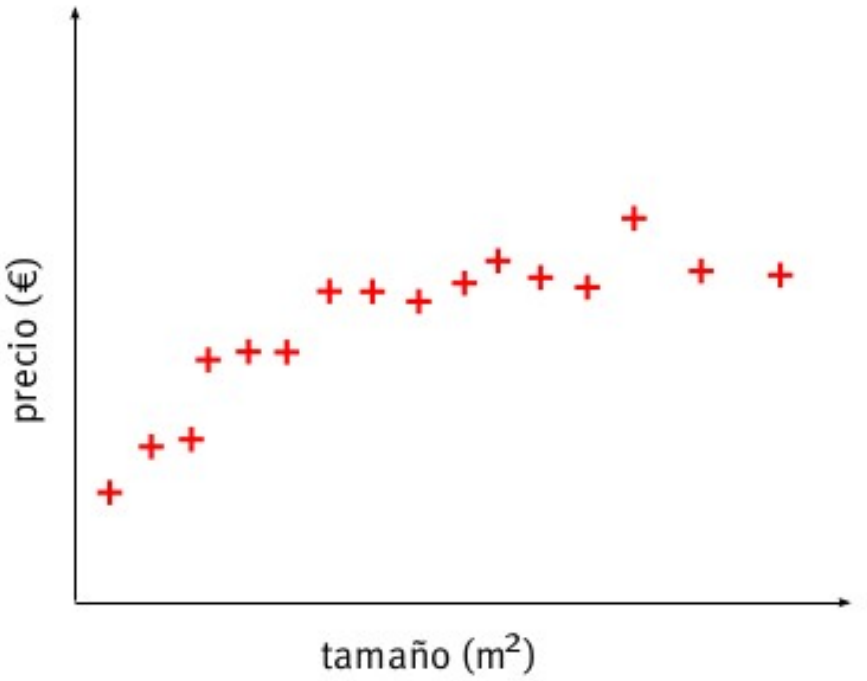


Figura: Ejemplo de problema de aprendizaje supervisado de regresión:
Dados estos datos, un amigo tiene una casa de 75 metros cuadrados,
¿por cuánto podría esperar venderla?

Clasificación

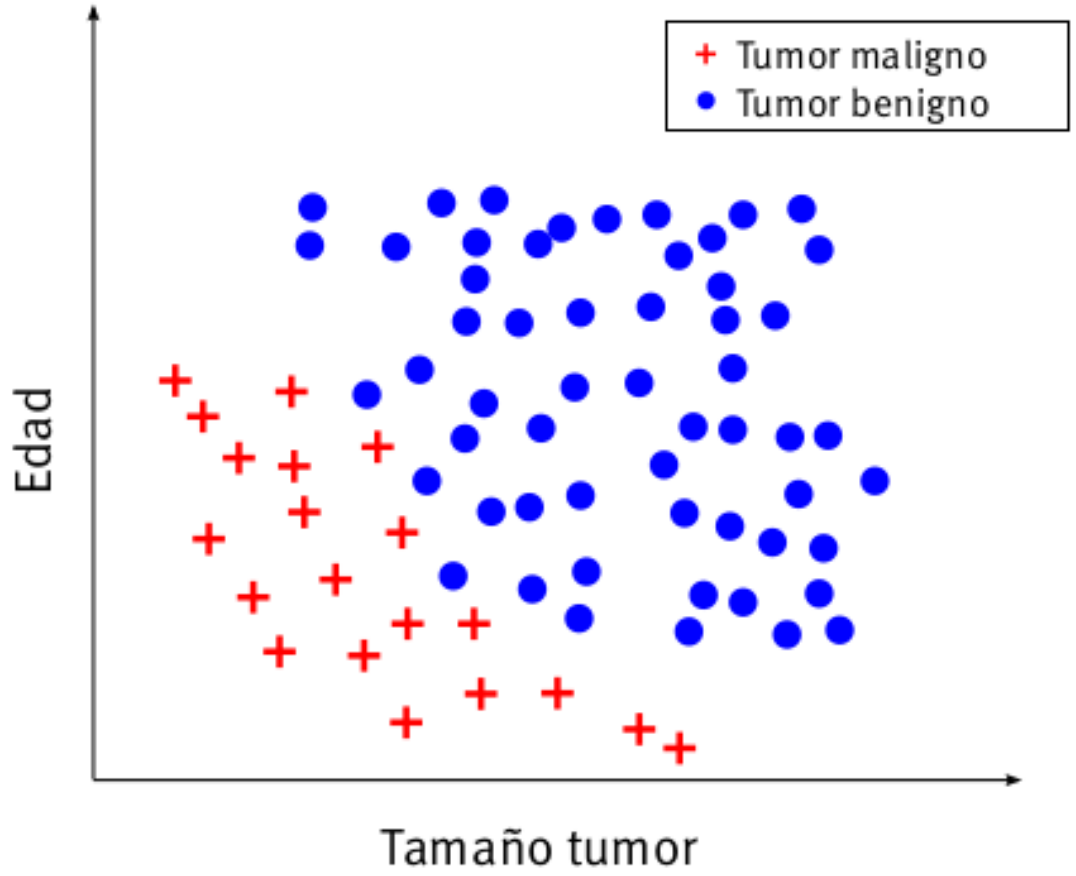


Figura: Ejemplo de problema de clasificación ¿Podrías estimar un
diagnóstico basado en el tamaño del tumor y la edad del paciente?

Clasificación

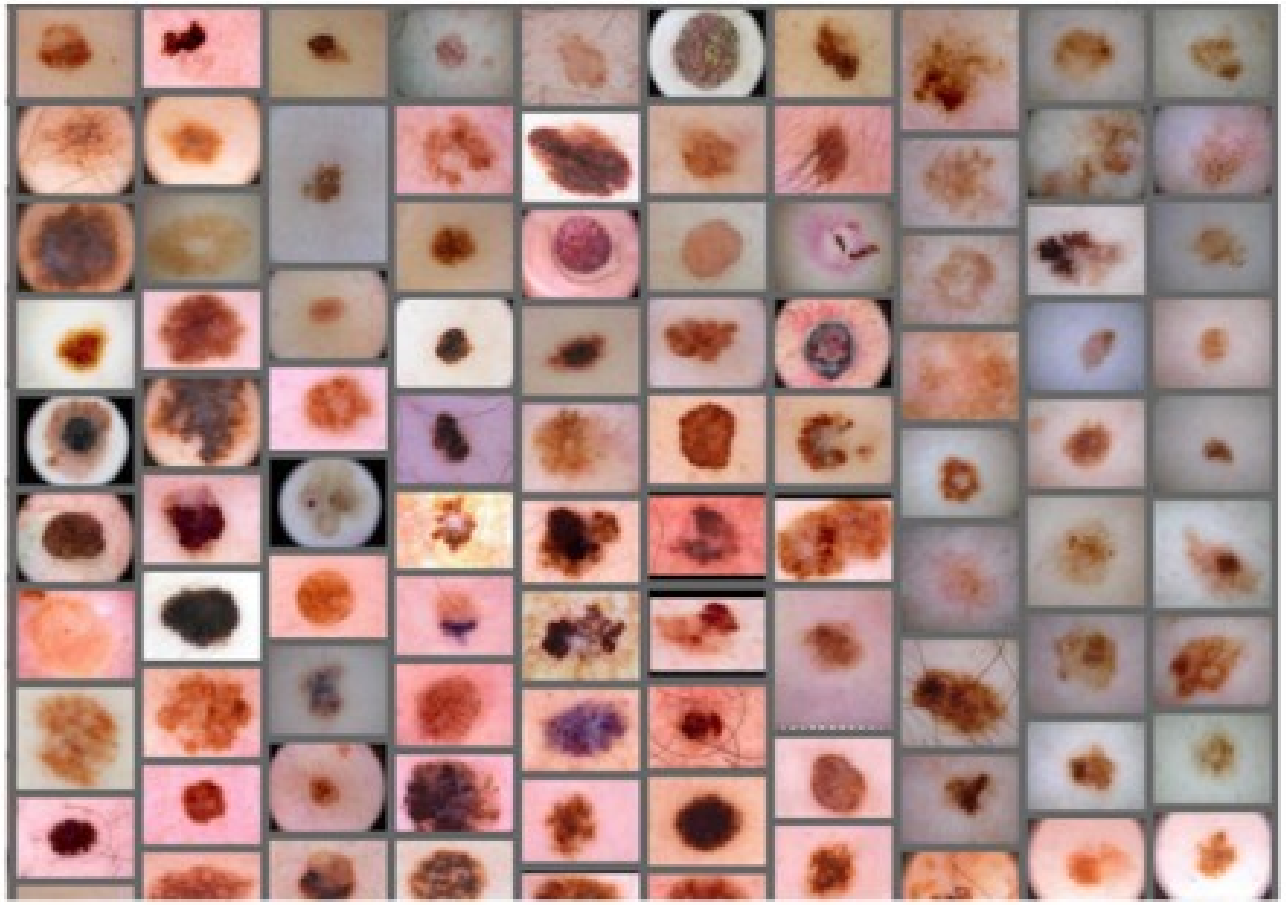


Figura: Ejemplo de problema de clasificación de imágenes. ¿Data una nueva fotografía podemos decir que es un cáncer de piel en base a nuestro modelo?

Predicción: reglas de asociación

Figura: ¿Qué productos suelen ir juntos en las cestas de la compra?¿Qué probabilidad hay de que una persona que compre el producto A compre el producto B?

GRACIAS