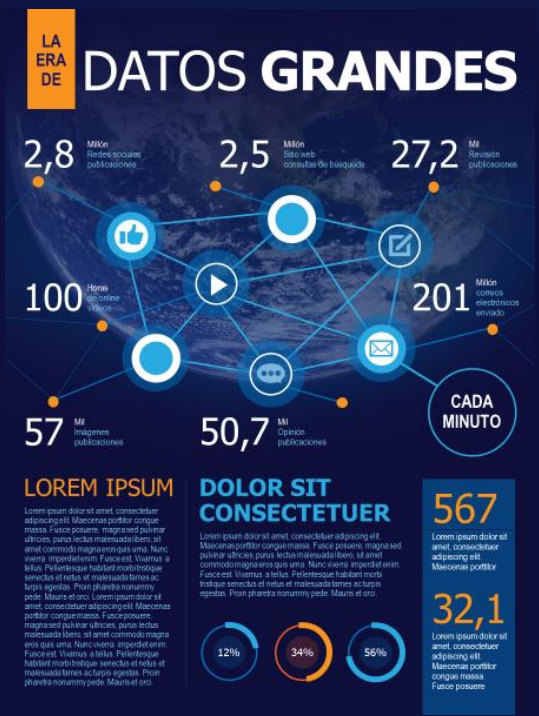
BIG DATA


LA ERA DE
DATOS GRANDES
2,8 Millón Redes sociales publicaciones
2,5 Millón Sitio web consultas de búsqueda
27,2 Mil Revisión publicaciones
100 Horas
201 Millón correos electrónicos enviado
CADA MINUTO
57 Mil Imágenes publicaciones
50,7 Mi Opinión publicaciones
LOREM IPSUM
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci.
DOLOR SIT CONSECTETUER
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna. Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci.
12%
34%
56%
567
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor
32,1
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere

BIG DATA

El big data es el análisis de un gran volumen de datos. El big data se caracteriza por las tres V:

Volumen: la cantidad de datos importa. . Pueden ser datos no estructurados de valor desconocido como feeds de datos de Twitter, cliccs de páginas web sensores, …

Velocidad: es el ritmo al que se reciben los datos y al que se aplica alguna acción.

Variedad: hace referencia a los diversos tipos de datos disponibles. Los datos convencionales eran estructurados –seguían un orden-, ahora los datos pueden ser no estructurados o semiestructurados, como un texto, audio o video.

Historia del Big Data: antecedentes

18000 AEC. En el Paleolítico Superior se empleaban **rudimentarios métodos de almacenamiento de datos** con el empleo de palos o muescas en huesos.

2400 AC. En Babilonia se extiende el uso del **ábaco**, un sistema para realizar cálculos. En esta época surgen también las primeras bibliotecas como lugares para almacenar y consultar conocimiento.

48 AC. Los Romanos invaden Alejandría y accidentalmente destruyen su famosa biblioteca. Parte de los fondos se trasladaron a otros lugares, pero la mayoría de la colección fue quemada, perdida o robada.

1663. John Graunt realiza el primer experimento de **análisis de datos** estadísticos conocido. Con los datos de defunciones, teoriza un sistema de alerta para la peste bubónica en toda Europa.

1792. Aunque hay constancia de análisis estadísticos desde las Guerras del Peloponeso y la palabra *estadística* se acuña en Alemania unos años antes; en 1792 se asocia el término a la "**colección y clasificación de datos**".

1865. Aparece por primera vez el término **business intelligence**, en la enciclopedia comercial de Richard Millar Devens. En ella describe cómo el banquero Henry Furnese logró una importante ventaja competitiva recogiendo, estructurando y analizando datos clave de su actividad. La inteligencia de negocio es sin duda uno de los grandes motores de la analítica dentro de la historia del big data.

1880. Herman Hollerith, empleado del censo estadounidense, desarrolla su **máquina tabuladora**. Con ella consigue reducir un trabajo de 10 años a 3 meses. Este ingeniero funda una compañía que posteriormente se conocería como IBM.

1926. Nikola Tesla predice la tecnología inalámbrica. Según su visión, el planeta en un gran cerebro en el que todo está conectado, por lo que deberíamos ser capaces simplificar el uso del **teléfono**. Predice que cada hombre llevará uno en su propio bolsillo.

1928. El ingeniero alemán Fritz Pfleumer patenta el **primer sistema** magnético para almacenar datos. Sus principios de funcionamiento se utilizan hoy en día.

1944. Primer intento de **conocer la cantidad información** que se crea. Se trata de un estudio académico de Fremont Rider, que pronostica 200 millones de libros en la Universidad de Yale en 2040, almacenados 6.000 millas de estanterías.

1958. El informático alemán Hans Peter Luhn, define la **inteligencia de negocio**: la habilidad de percibir las interrelaciones de los hechos presentados para guiar acciones hacia un objetivo deseado. En 1941 pasó a ser Gerente de Recuperación de Información en IBM.

1962. Se presenta IBM Shoebox en la Expo de 1962. Creada por William C. Dersch supone el primer paso en el **reconocimiento de voz**, capaz de registrar palabras en inglés en formato digital.

Historia del Big Data

1965. Se proyecta el **primer *data center*** en Estados Unidos, para guardar documentación de impuestos y huellas dactilares en cintas magnéticas. Un año antes comienzan a surgir voces que alertan del problema de guardar la ingente cantidad de datos generada.

1970. IBM desarrolla el **modelo relacional de base de datos**, gracias al matemático Edgar F. Codd. Este científico inglés es también responsable de las doce leyes del procesamiento analítico informático y acuño el término OLAP.

1976. Se populariza el uso de MRP (software de gestión de materiales), antecedentes de los **ERP** actuales, que mejoran la eficiencia de las operaciones en la empresa; además de generar, almacenar y distribuir datos en toda la organización.

1989. Erik Larson habla por primera vez de **Big Data** en el sentido que conocemos la expresión hoy en día. La revista Harpers Magazine recoge su artículo, en el que especula sobre el origen del correo basura que recibe. En torno a este año se empiezan a popularizar las herramientas de business intelligence para analizar la actividad comercial y el rendimiento de las operaciones.

1991. **Nace internet (Tim Berners-Lee)**, la gran revolución de la recolección, almacenamiento y análisis de datos.

1996. Los precios del **almacenamiento de datos** empiezan a ser accesibles con un coste eficiente.

1997. **Google** lanza su sistema de búsqueda.

1999. El término Big Data es analizado por primera vez en un **estudio académico**.

2001. Doug Laney, de Gartner, define **las 3 V's del Big Data**.

2005. Nace la Web 2.0, una web donde predomina el contenido creado por los usuarios. Hadoop=Big Data (libre)

2007. La revista Wired publica un artículo que lleva el concepto de **Big Data a las masas**.

2010. Los datos que se generan en dos días equivalen a la **cantidad de datos generados** desde el inicio de la civilización hasta 2003, según Eric Schmidt (Google).

2013. El archivo de mensajes públicos de **Twitter** en la Biblioteca del Congreso de Estados Unidos llega a los 170 billones de mensajes, creciendo a ritmo de 500 millones al día..

2014. Los **móviles** superan a los ordenadores en accesos a internet. La conexión casi continua contribuye a generar muchos más datos y mejora la conectividad con otros dispositivos.

2016. El Big Data se convierte en la **palabra de moda**. Se generaliza la contratación de expertos en Big Data, el Machine Learning llega a las fábricas y el Internet de las Cosas empieza a impregnarlo todo.

2017. Los datos llegan a **las masas**. La gente controla sus patrones de descanso con pulseras, sabe en qué se gasta el dinero con aplicaciones móviles y se informa sobre la posesión de balón de su equipo de fútbol. Los datos están en todas partes y la población está ya predispuesta a usarlos.

Futuro. ¿Qué nos deparará el futuro? Muy difícil de pronosticar, pero seguramente un aumento de datos y la consiguiente necesidad de tecnología para recogerlos, adaptarlos, almacenarlos y analizarlos. La **computación cuántica** está a la vuelta de la esquina y la historia del big data sigue avanzando.

Big data vs Ciencia de datos

Data Science o Ciencia de datos es el campo que abarca la limpieza, preparación y análisis de los datos, a través de las Matemáticas, la Estadística y otras herramientas informáticas para extraer conocimiento de los datos. Incluye Big Data.

- Reúne datos de múltiples disciplina y los compila.
- Machine learning para análisis predictivo, Deep learning, inteligencia artificial (Rijmenam,2013).
- Extrae información útil.

Big Data es una gran colección de conjuntos de datos que no se pueden almacenar en un sistema tradicional. Su tamaño puede variar hasta peta-bytes.(10^15 = mil billones; tera= 10^12, giga= 10^9; mega=10^6)

- Captura de datos mediante herramientas, metodologías o tecnologías.
- Almacena datos a través de plataformas
- Busca datos
- Analiza datos

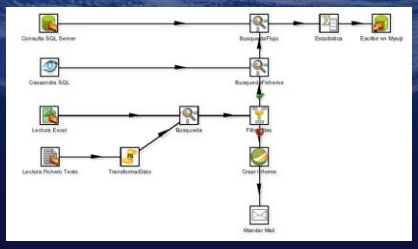
Usos o aplicaciones:

Ciencia de datos: utiliza búsquedas en internet y motores de búsqueda para entregar mejores resultados. Marketing.

Big Data: (servicios financieros, tarjetas de crédito, bancos minoristas, compañías de seguros) Analiza al cliente, su cumplimiento, el fraude y el análisis operacional. ARMAS DE DESTRUCCIÓN MATEMÁTICA (Cathy O'Neil-2018)

Ejemplo de Big Data

- Tenemos desplegados 1000 sensores y necesitamos inserts y updates rápidos. Buscamos escalabilidad en una Base de Datos para llegar a desplegar 100.000 sensores; con una Base de Datos que escale en nº de nodos y así mantener su rendimiento.
- Disponemos de dato:
  - BD relacionales: Tablas de EXCEL
  - BD no relacionales: Imágenes en Cassandra.
- Extracción de datos: Kettle (Pentaho Data Integration (PDI)), para extraer informes, etc.



TIPOS DE BASES NO RELACIONALES

- BD Key-Value: con 2 columnas, la clave y un nº binario correspondientes a videos, imágenes, texto,….
- BD orientadas a documentos: =, guardado en JSON como MongoDB

- BD orientadas a grafos: se almacena en esquema de grafos
- BD orientadas a objetos: pueden tener “herencia” entre objetos.

Amazon y Big Data

- Amazon S3: un disco duro en internet. Hasta 5 terabytes, se necesita registro y una tarjeta de crédito para que en caso de pasarse, cobren.

"Como parte de la Capa de uso gratuito de AWS (Amazon Web Services)podrá empezar gratis con Amazon S3. Al registrarse, los clientes nuevos de AWS reciben 5 GB de almacenamiento estándar en Amazon S3, 20 000 solicitudes GET, 2 000 solicitudes PUT y 15 GB de

- transferencia de datos saliente al mes durante un año."

- 1.- Creamos un "Bucket": s3://BUCKET/CARPETA/ARCHIVO
  - Cuando los datos se comparten en AWS, cualquiera puede analizarlos y crear servicios sobre ellos utilizando una amplia gama de productos informáticos y de análisis de datos, incluidos Amazon EC2 , Amazon Athena , AWS Lambda y Amazon EMR . Compartir datos en la nube permite a los usuarios de datos dedicar más tiempo al análisis de datos en lugar de la adquisición de datos.
- Amazon EC2: obtener servidores virtuales, configurar la seguridad y las redes y administrar el almacenamiento.
- Amazon Athena: consultas en SQL (pago)
- Aws Lambda: ejecuta código como respuesta a eventos y escala automáticamente.
- Amazon EMR: plataforma de big data se combinan con la escalabilidad dinámica de Amazon EC2 y el almacenamiento escalable de Amazon S3.

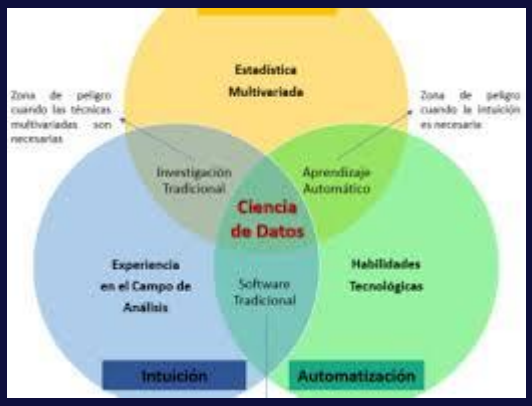
Amazon Elastic MapReduce (EMR) permite centrar los esfuerzos en el procesamiento o analítica de datos sin que un cliente tenga que preocuparse en crear y configurar esa estructura BigData. Ya lo hacen por el cliente.

- CÓMO FUNCIONA
- Amazon ha creado un formulario el cual permite configurar un cluster BigData. EMR utiliza Apache Hadoop como motor de procesamiento distribuido. Hadoop es un framework de software Java de código abierto que permite utilizar aplicaciones de uso intensivo de datos que se ejecutan en agrupaciones de equipos sencillos de gran tamaño. Hadoop implementa un modelo informático denominado "MapReduce" que divide el trabajo en pequeños fragmentos, cada uno de los cuales puede ejecutarse en cualquiera de los nodos que forman la agrupación de equipos.

Ciencia de datos

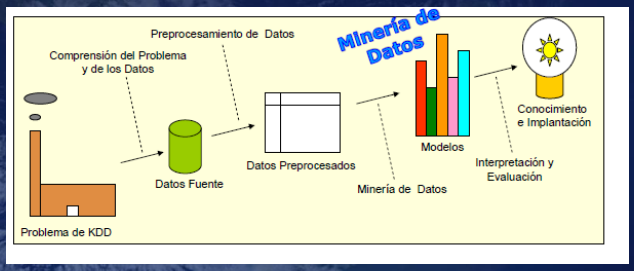
- La Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas a la extracción de conocimiento de datos, incluyendo Big Data

- Se requiere dominar las ciencias matemáticas y la estadística, conocimientos de programación (Phyton y otros lenguajes), ciencias de la computación y analítica.



Ciencia de datos

- Minería de Datos: descubrimiento de patrones interesantes en una base de datos (usualmente grande)
- Proceso de KDD (Knowledge Discovery from Databases): limpieza, integración, reducción de datos, transformación, minería de datos, evaluación y presentación del conocimiento.
- Técnica de Minería de Datos:
  - Clasificación
  - Regresión
  - Agrupamiento o Clustering.
  - Asociación
  - Tendencias



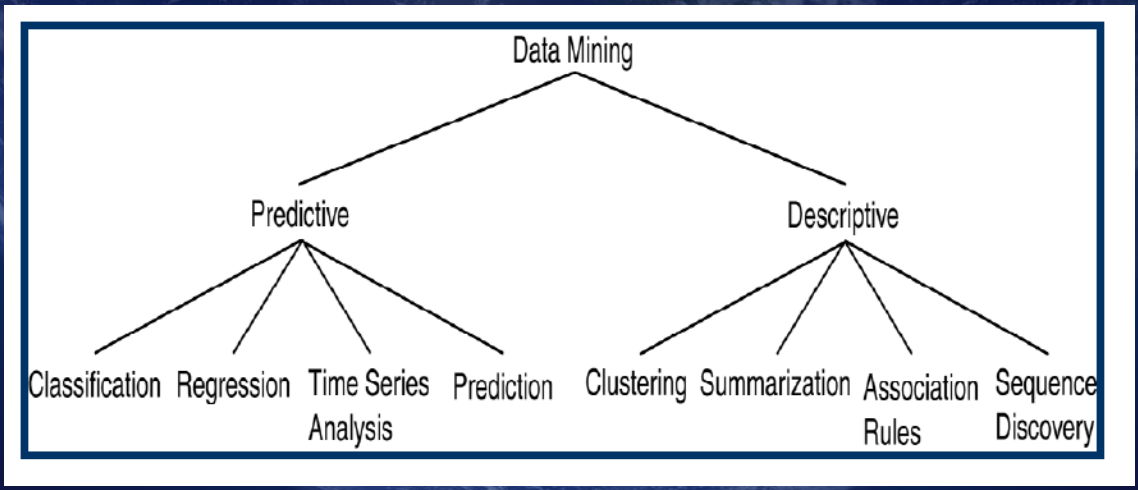
- Machine learning (Kaggle: web para practicar en la resolución de problemas reales y adquirir habilidades en Ciencia de Datos) Phyton: lambda, reduce, filter y map.
- Herramientas: Knime (**Konstanz Information Miner) plataforma de minería de datos que permite el desarrollo en un entorno visual. *KEEL*** (siglas de Knowledge Extraction based on Evolutionary Learning -Extracción de Conocimiento basado en Aprendizaje Evolutivo-) es un conjunto de herramientas de software de aprendizaje automático, desarrollados bajo el proyecto granadino nacional TIC2002-04036-C05, TIN2005-08386-C05 y TIN2008-06681-C06.

http: www.kdnuggets.com: web sobre inteligencia artificial, análisis, big data, minería de datos, ciencia de datos y aprendizaje automático .

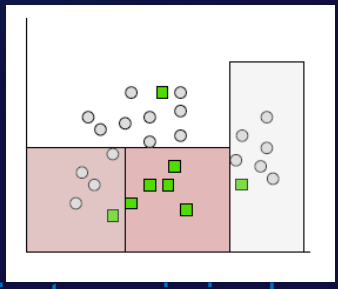
- El poder de los datos: asociaciones, salud, transacciones, recomendación (Amazon), Twitter y salud, tweets y campañas electorales.

Técnicas de Minería de Datos

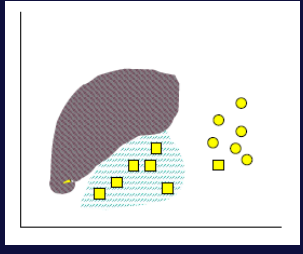
- Métodos predictivos: se utilizan algunas variables para predecir valores desconocidos de otras variables.
- Métodos descriptivos: encuentran patrones interpretables que describen los datos.



- Aprendizaje Supervisado: aprende a partir de un conjunto de instancias pre-etiquetadas. Predice.

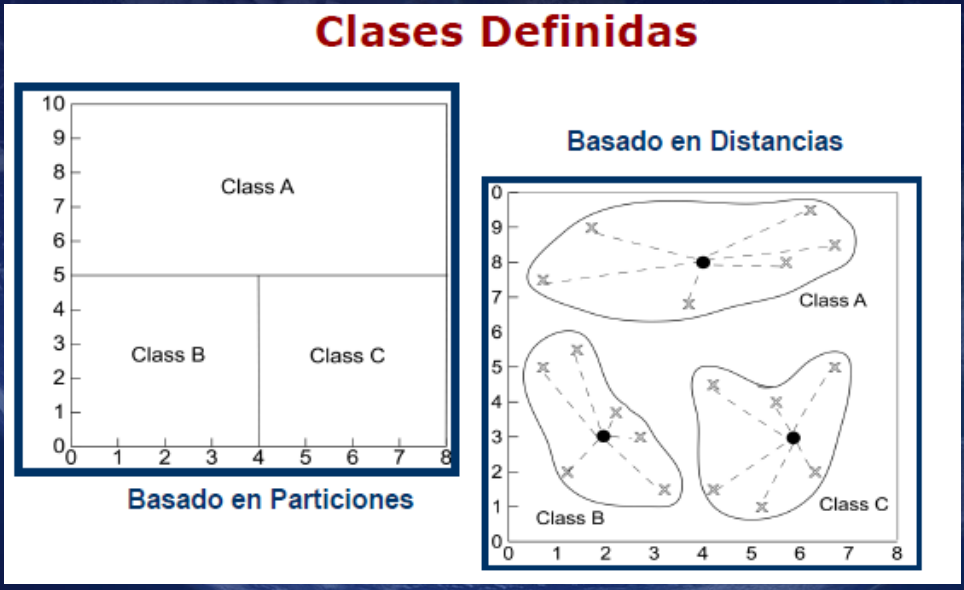


- Aprendizaje no supervisado: no hay conocimiento a priori sobre el problema, no hay instancias etiquetadas, no hay supervisión sobre el procedimiento. Clustering. Describe

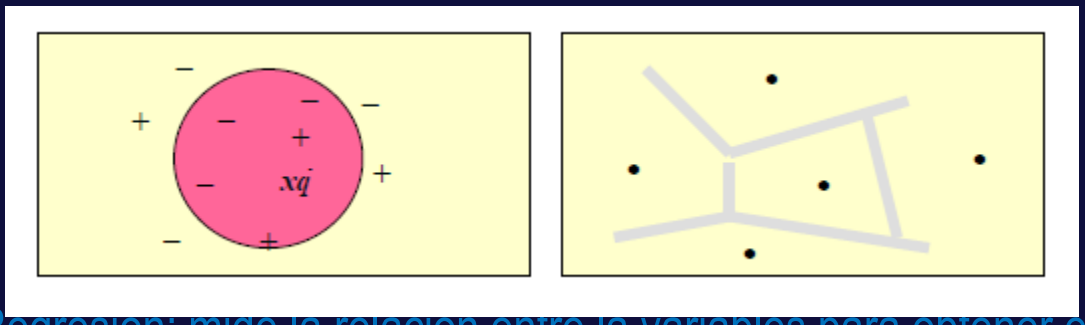


Métodos predictivos

- Clasificación:



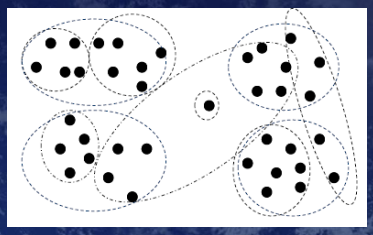
  - Modelos Interpretables: árboles de decisión, listas de decisión.
  - Modelos no interpretables: Clasificadores basados en casos (k-NN). Redes neuronales-DeepArt, redes bayesianas. SVMs (Suppot Vector Machines)
    - K-NN devuelve la clase más repetida de entre todos los k ejemplos de entrenamiento cercanos a xq.
    - Diagrama de Voronoi: superficie de decisión inducida por 1-NN para un conjunto dado de ejemplos de entrenamiento.

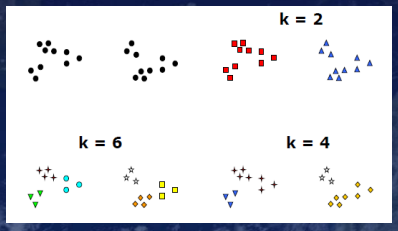


- Regresión: mide la relación entre la variables para obtener el valor de la variable control.
- Series temporales
- Predicción.

Métodos descriptivos.

- Agrupamientos o clustering.:





- marketing por tipos de clientes en la BD. Identificación de cultivos con BD y observación. Seguros con productos. Planificación urbana con identificación de inmuebles. WWW con clasificación de documentos, ficheros, accesos,..

- Resumen de datos.
- Reglas de asociación.
- Descubrimiento de asociaciones.

Práctica

* Big Data Bayesiano

* Clústering y dependencia entre datos:
    PAST: Programa estadístico