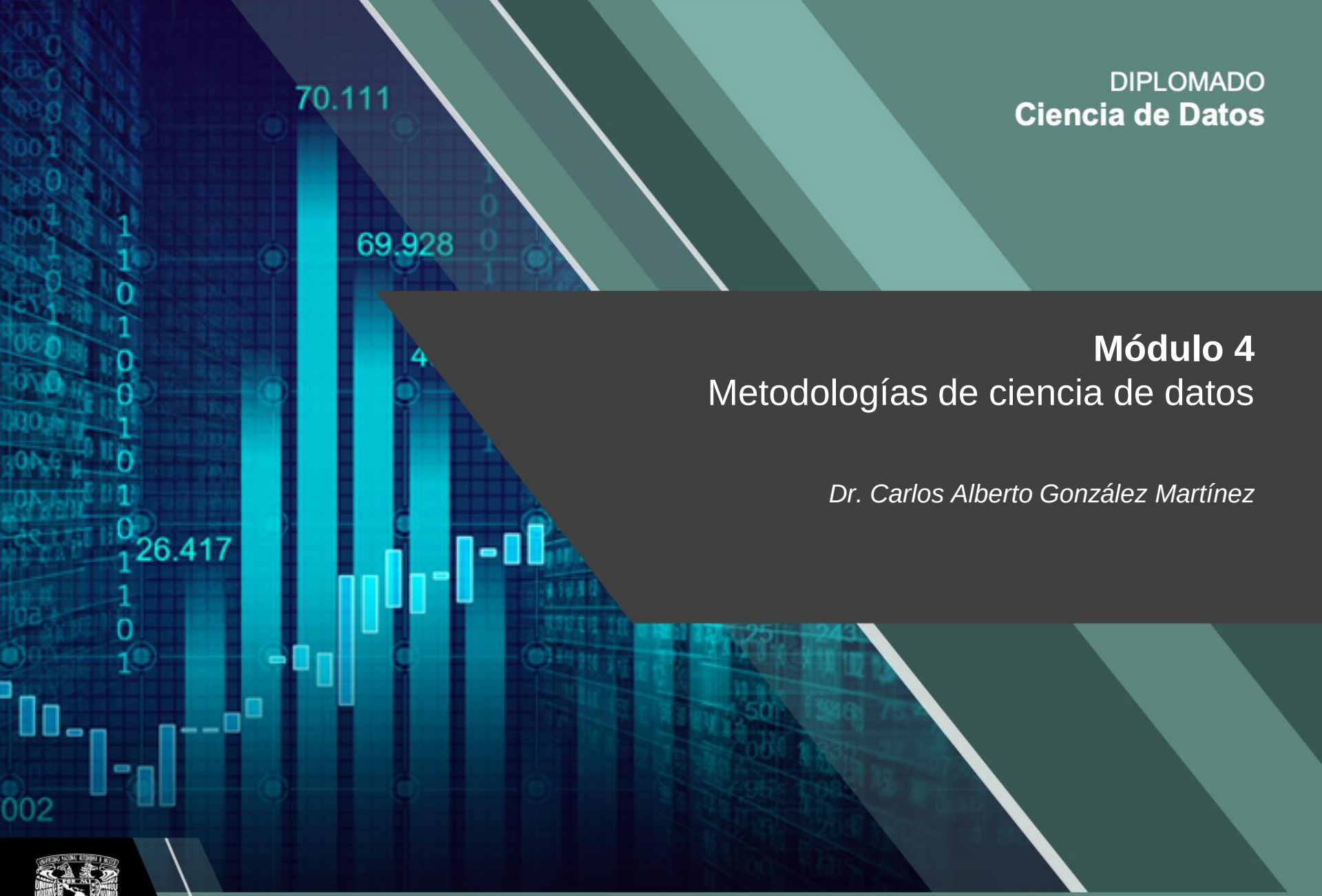


DIPLOMADO
Ciencia de Datos

Módulo 4
Metodologías de ciencia de datos

Dr. Carlos Alberto González Martínez

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

INTRODUCCIÓN: MINERÍA DE DATOS

Objetivo

El participante identificará las características principales de la minería de datos y el volumen de datos, así como las principales herramientas comerciales y de *open source* para su manejo.

Minería de datos

Contenido

Introducción a la minería de datos

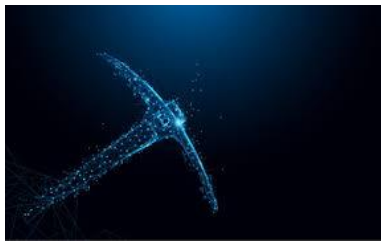
1. Volumen de datos
2. Qué es la minería de datos
3. Herramientas de la minería de datos
4. Herramientas comerciales

Introducción a la minería de datos

Desde los años sesenta los estadísticos manejaban términos como **data fishing**, **data mining** o **data archaeology**, con la idea de encontrar correlaciones sin una hipótesis previa, en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, entre otros, comenzaron a consolidar los términos de **data mining**.

A finales de los años ochenta, sólo existían dos empresas dedicadas a esta tecnología. En el 2002 existían más de 100 empresas en el mundo, que ofrecían alrededor de 300 soluciones.

El data mining es una tecnología compuesta por etapas, que integra varias áreas, que no se debe confundir con un gran software.



Introducción a la minería de datos

A fines de los 80 apareció un nuevo campo de investigación llamado **KDD** (Knowledge Discovery in Databases)

KDD es el proceso no trivial de identificar patrones a partir de los datos con las siguientes características:

- Válidos
- Novedosos
- Potencialmente útiles
- Comprensibles

La revolución digital ha permitido que la captura de datos sea fácil, y su almacenamiento tenga un costo casi nulo.

Minería de datos

Contenido

Introducción a la minería de datos

1. **Volumen de datos**
2. Qué es la minería de datos
3. Herramientas de la minería de datos
4. Herramientas comerciales

1. Volumen de datos

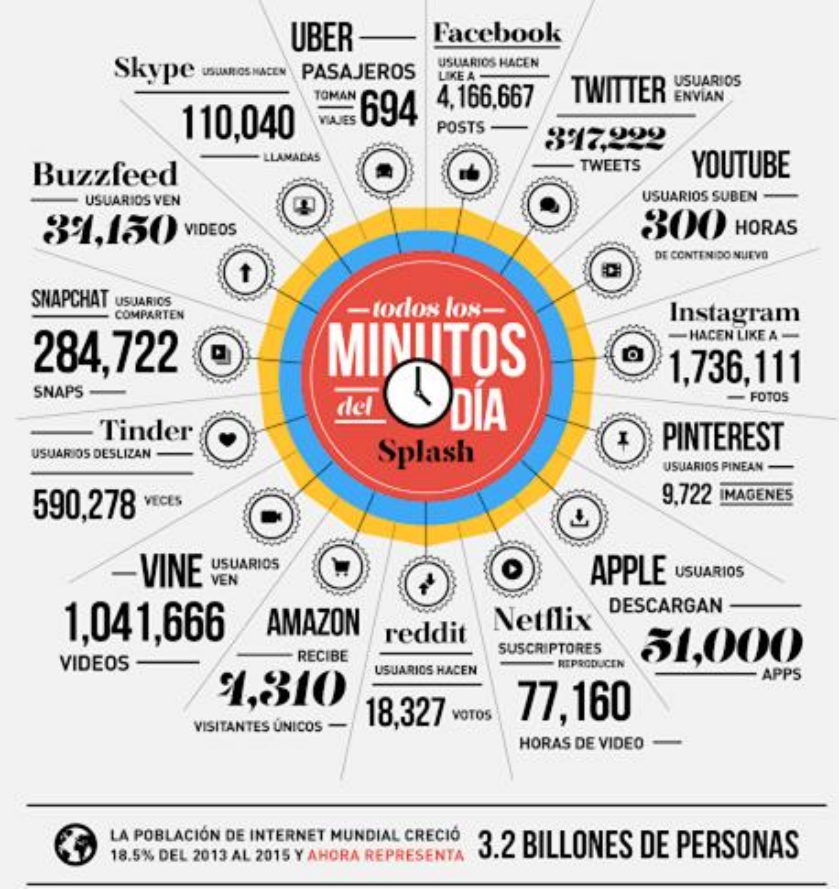
Enormes cantidades de datos son recogidas y almacenadas en BD en la vida diaria.

1. Volumen de datos

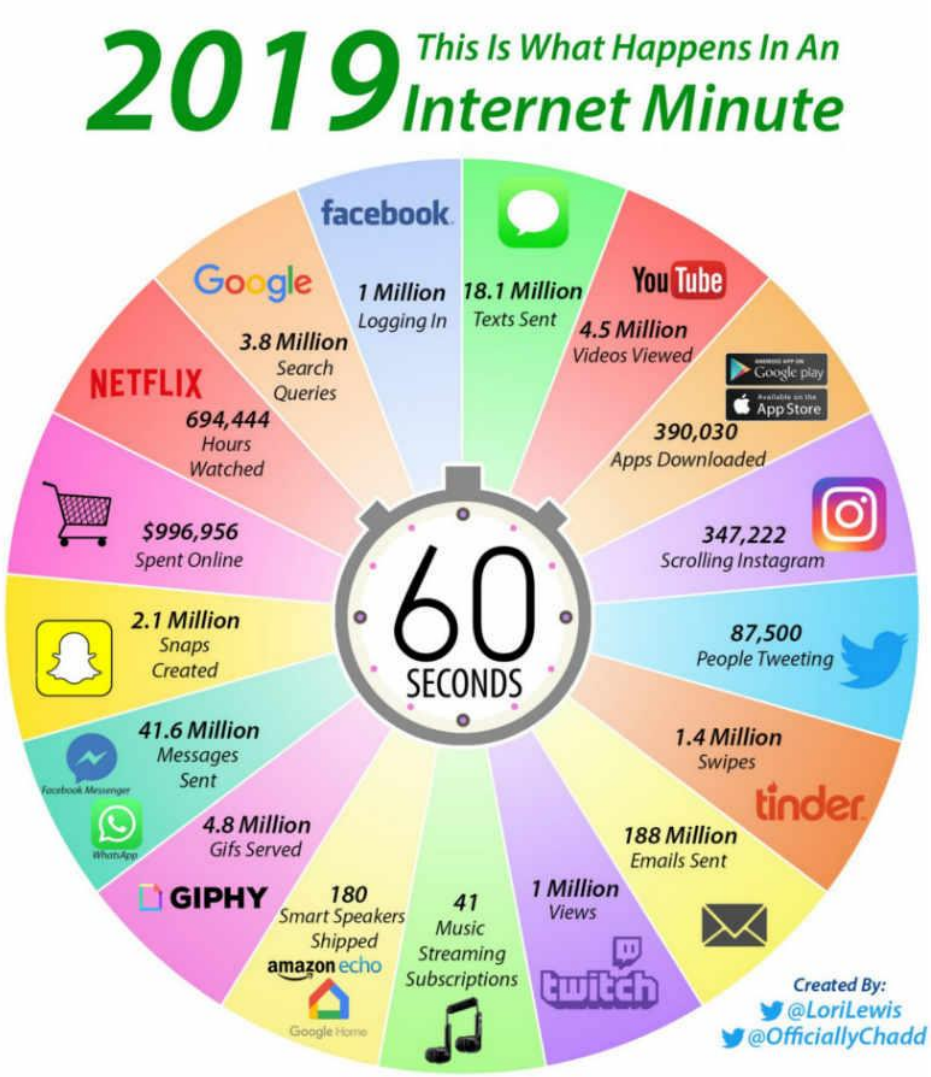
Una tendencia actual es que todo genere datos para ser analizados.

1. Volumen de datos

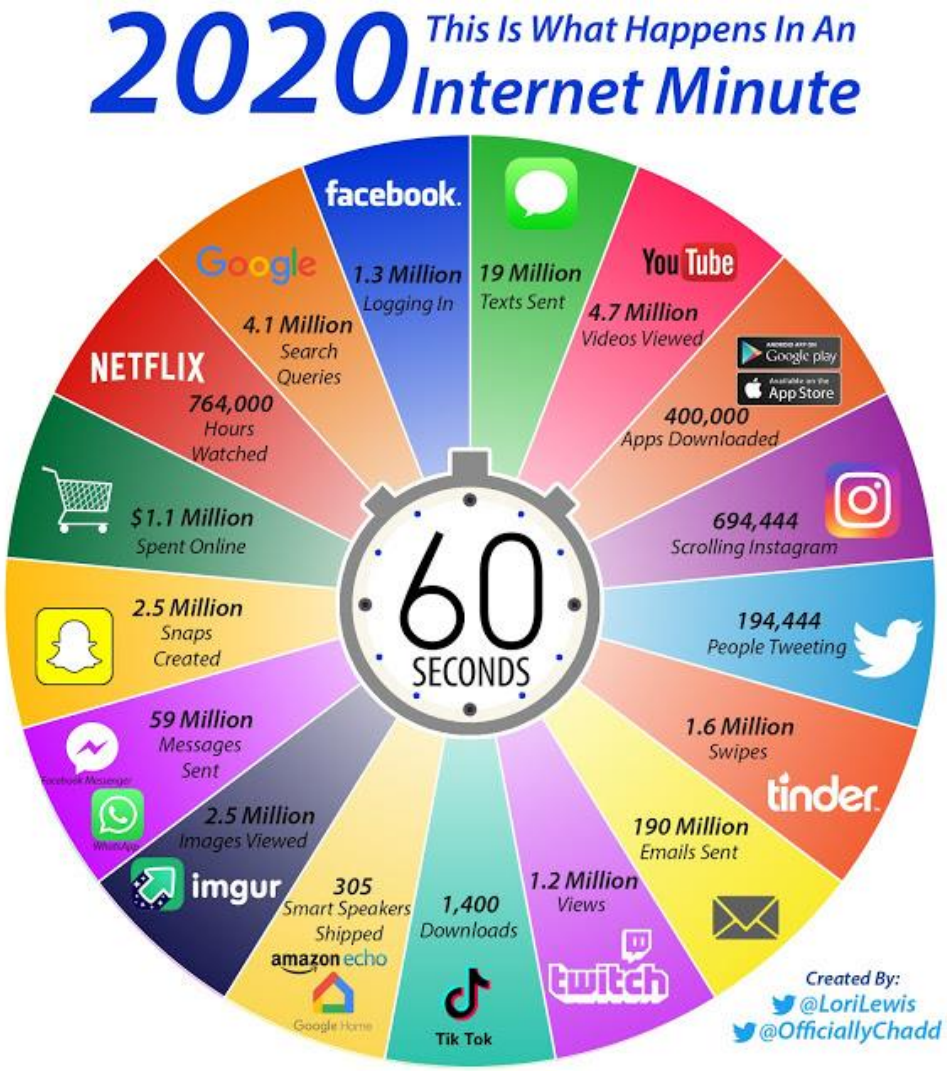
Una infografía sobre cuántos datos se generan en Internet cada minuto.



1. Volumen de datos



1. Volumen de datos



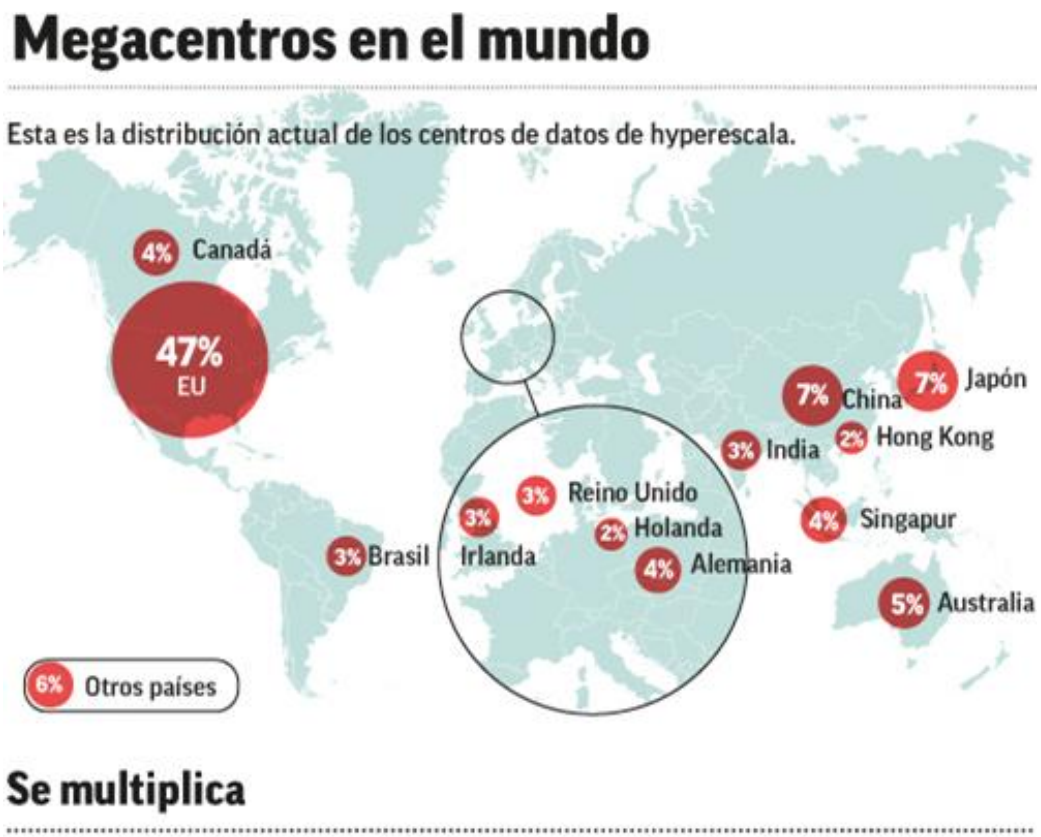
1. Volumen de datos

Para 2020 Cisco estimó que la información se triplicará en sus centros de datos

Más empresas impulsando Internet de las Cosas se conectan para aprovechar, por ejemplo, el análisis con 'big data'. Actualmente circulan 6.5 zettabytes o 6 mil 979 millones 321 mil 856 terabytes en los centros de datos; pero para 2020, Cisco estima que el tráfico llegará a 15.3 zettabytes de datos.

De acuerdo a Rodolfo Molina, director senior de cloud (la nube) y gestión de servicios de **Cisco** para Latam, actualmente un avión 787 de Aeroméxico produce 40 terabytes al día.

1. Volumen de datos



1. Volumen de datos

Estas son las previsiones de tráfico de los centros de datos por región de Cisco.

	2015	2020
Norteamérica	**2.2** zettabytes	**7.1** zettabytes
Europa Occidental	**843** exabytes	**2.7** zettabytes
Europa del Este y Central	**191** exabytes	**632** exabytes
Latinoamérica	**191** exabytes	**533** exabytes
Medio Oriente y África	**105** exabytes	**451** exabytes
Asia Pacífico	**1.2** zettabytes	**4.0** zettabytes

¿A qué equivale un zettabyte?

La información esperada en centros de datos se miden en zettabytes.

1 KILOBYTE	→	1024 BYTES
1 GIGABYTE	→	1024 KILOBYTES
1 TERABYTE	→	1024 GIGABYTES
1 PETABYTE	→	1024 TERABYTES
1 EXABYTE	→	1024 PETABYTES
1 ZETTABYTE	→	1024 EXABYTES
1 YOTTABYTE	→	1024 ZETABYTES

Minería de datos

Contenido

Introducción a la minería de datos

1. Volumen de datos
2. **Qué es la minería de datos**
3. Herramientas de la minería de datos
4. Herramientas comerciales

2. Qué es minería de datos

Para analizar enormes cantidades de datos, las herramientas tradicionales de gestión de datos y las herramientas estadísticas no son adecuadas.

Los sistemas tradicionales de explotación de datos están basados en la existencia de hipótesis o modelos previos.

La Minería de Datos busca el descubrimiento del conocimiento **sin una hipótesis** preconcebida.

La minería de datos tiene como propósito la identificación de un conocimiento obtenido, a partir de las bases de datos que aporten hacia la toma de una decisión.

2. Qué es minería de datos

- Minería de datos es la exploración y el análisis de grandes cantidades de datos, con el objeto de encontrar patrones y reglas significativas (conocimiento).

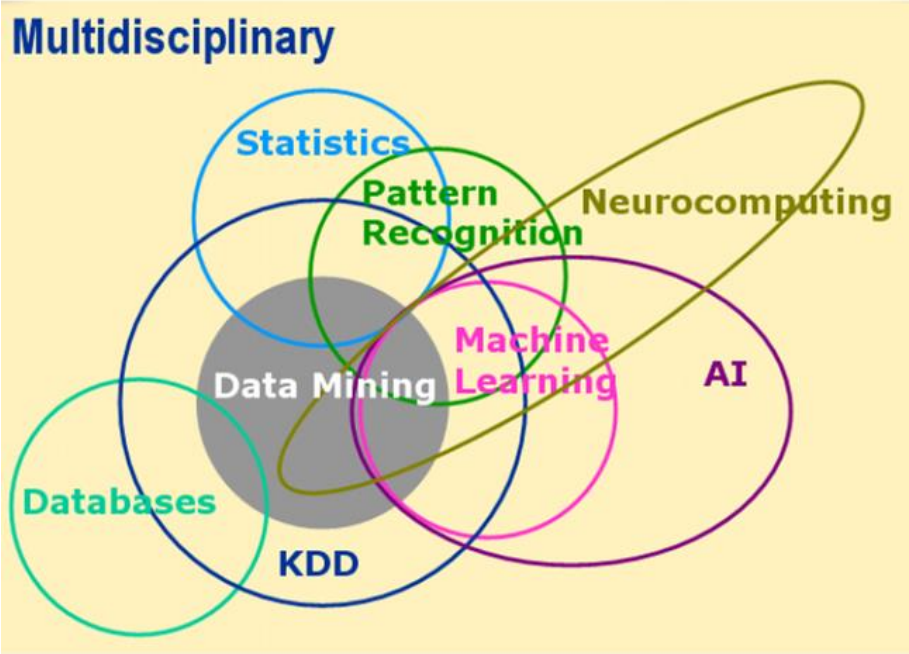
- Es un mecanismo de explotación, que consiste en la búsqueda de información valiosa en grandes volúmenes de datos.

- Análisis de grandes volúmenes de datos para encontrar relaciones no triviales, y para resumirlos de manera que sean entendibles y útiles. (**Hand, Mannila y Smyth).**

- Proceso de extracción de conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten and Frank, 2000).

2. Qué es minería de datos

No existe un solo enfoque para minería de datos, sino un conjunto de técnicas que se pueden utilizar de manera independiente o en combinación.



2. Qué es minería de datos

Entendemos por ***data mining*** el proceso de analizar datos provenientes de distintas fuentes de información, con el objetivo de extraer información y conocimiento útil.

- Por **útil** entendemos que debe haber una alineación con unos objetivos previamente establecidos, tanto desde el punto de vista del negocio, como desde el punto de vista más concreto del data mining.

- Por **información** entendemos asociaciones, relaciones y estadísticas básicas, capaces de responder a preguntas como qué productos se están vendiendo, y cuándo y dónde se está produciendo esta venta.

- Por **conocimiento** *(knowledge)* entendemos patrones históricos, basados en la observación del pasado, y también tendencias futuras, basadas en técnicas predictivas.

Minería de datos

Contenido

Introducción a la minería de datos

1. Volumen de datos
2. Qué es la minería de datos
3. **Herramientas de la minería de datos**
4. Herramientas comerciales

3. Herramientas de minería de datos

Entre las herramientas de minería de datos, las de software libre cobran una gran importancia, ya que nos permiten llevar a cabo procesos de asociación, selección, conglomerados, etcétera, con un gran volumen de datos, haciendo una mínima inversión. Entre las herramientas líderes en este ramo, tenemos las siguientes (referencia):

- Orange
- Rapidminer
- Weka
- JHepWork
- KNIME

3. Herramientas de minería de datos

Orange es un software de código abierto. Es una suite de software para minería de base de datos y aprendizaje automático, basado en componentes que cuentan con un fácil y potente, rápido y versátil, *front-end* de programación visual para el análisis exploratorio de datos y visualización, y librerías para Python y secuencias de comando.

Contiene un completo juego de componentes para preprocesamiento de datos, característica de puntuación y filtrado, modelado, evaluación del modelo y técnicas de exploración. Está escrito en C++ y Python, y su interfaz gráfica de usuario se basa en la plataforma cruzada del frameworkQt.

3. Herramientas de minería de datos

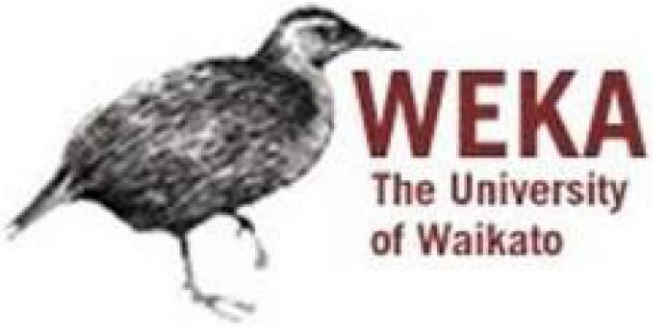
RapidMiner, antes llamado YALE, es un ambiente de experimentos en aprendizaje automático y minería de datos, que se utiliza para tareas de minería de datos, tanto en investigación como en el mundo real.

Permite a los experimentos estar compuestos de un gran número de operadores anidables arbitrariamente, que se detallan en archivos XML y se hacen con la interfaz gráfica de usuario de RapidMiner.

RapidMiner ofrece más de 500 operadores para todos los principales procedimientos de máquina de aprendizaje, y también combina esquemas de aprendizaje y evaluadores de atributos del entorno de aprendizaje Weka.

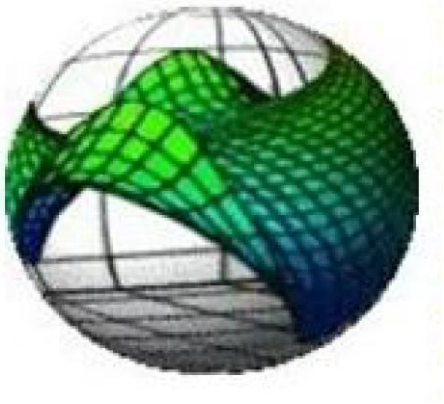
Está disponible como una herramienta *stand-alone*, para el análisis de datos y como motor para minería de datos, que puede integrarse en tus propios productos.

3. Herramientas de minería de datos



Escrito en Java, Weka (Entorno Waikato para el Análisis del Conocimiento) es una conocida suite de software para máquinas de aprendizaje, que soporta varias tareas típicas de minería de datos, especialmente preprocesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección. Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano o relación, donde cada punto marcado es etiquetado por un número fijo de atributos. WEKA proporciona acceso a bases de datos SQL, utilizando conectividad de bases de datos Java y puede procesar el resultado devuelto como una consulta de base de datos.

3. Herramientas de minería de datos



JHepWork. Diseñado para los científicos, ingenieros y estudiantes, jHepWork es un framework para el análisis de datos, libre y de código abierto, que fue creado como un intento de hacer un entorno de análisis de datos, usando paquetes de código abierto con una interfaz de usuario comprensible y para crear una herramienta competitiva a los programas comerciales. Esto se hace especialmente para las ploteos científicos interactivos en 2D y 3D y contiene bibliotecas científicas numéricas implementadas en Java, para funciones matemáticas, números aleatorios, y otros algoritmos de minería de datos. jHepWork se basa en Jython, un lenguaje de programación de alto nivel, pero codificación en Java. También puede ser usada para llamar librerías jHepWork numéricas y gráficas.

3. Herramientas de minería de datos

KNIME (Konstanz Information Miner) es una plataforma de código abierto de fácil uso y comprensible, para la integración de datos, procesamiento, análisis, y exploración.

Ofrece a los usuarios la capacidad de crear de forma visual flujos o tuberías de datos, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas.

KNIME está escrito en Java y está basado en Eclipse, y hace uso de sus métodos de extensión para soportar plugins, proporcionando así una funcionalidad adicional. A través de plugins, los usuarios pueden añadir módulos de texto, imagen, procesamiento de series de tiempo y la integración de varios proyectos de código abierto, tales como el lenguaje de programación R, WEKA, el kit de desarrollo de Química y LIBSVM.

Minería de datos

Contenido

Introducción a la minería de datos

1. Volumen de datos
2. Qué es la minería de datos
3. Herramientas de la minería de datos
4. **Herramientas comerciales**

4. Herramientas comerciales

Las herramientas comerciales de minería de datos cuentan con librerías predefinidas para el manejo de los datos, tanto en la conexión a las bases de datos, textos planos o archivos xls, la manipulación y la transformación de los mismos, como nodos específicos para llevar a cabo procesos de conglomerados, asociación, etcétera.

La presentación de los datos se puede hacer mediante un variado set de gráficas y los datos resultantes, ya sea en archivos o directamente a la base de datos.

Una de las mayores ventajas con las que cuentan las herramientas comerciales, es el seguimiento a los posibles bugs que se presenten en la versión, así como un soporte técnico formal, que va desde la instalación del software y la capacitación en el mismo, hasta el apoyo en el desarrollo de modelos.

4. Herramientas comerciales

Entre las herramientas comerciales líderes se encuentran:

- SAS
- SPSS Modeler

4. Herramientas comerciales

Lista detallada de precios	
Descripción de la pieza	**Precio de IBM, impuestos no incluidos USD**
IBM SPSS Modeler Professional Authorized User License + SW Subscription & Support 12 Months (D0EMZLL)	33,350.00
IBM SPSS Modeler Professional Authorized User Initial Fixed Term License + SW Subscription & Support 12 Months (D0EC5LL)	14,605.00
IBM SPSS Modeler Professional Concurrent User License + SW Subscription & Support 12 Months (D0EN7LL)	83,260.00
IBM SPSS Modeler Professional Concurrent User Initial Fixed Term License + SW Subscription & Support 12 Months (D0EC7LL)	36,570.00
IBM SPSS Modeler Premium Authorized User License + SW Subscription & Support 12 Months (D0EP4LL)	53,360.00
IBM SPSS Modeler Premium Authorized User Initial Fixed Term License + SW Subscription & Support 12 Months (D0ECFLL)	23,460.00
IBM SPSS Modeler Premium Concurrent User License + SW Subscription & Support 12 Months (D0EPGLL)	132,250.00

Referencias

Aguilar, J., (2020*). Introducción a Minería de Datos, Metodologías y Técnicas de Minería de datos*
. Recuperado de http://www.ing.ula.ve/~aguilar/actividad-docente/IN/transparencias/clase40.pdf

Domo, (2016). *Cuántos datos se generan en internet cada minuto.* Recuperado de https://ticsyformacion.com/2016/01/26/cuantos-datos-se-generan-en-internet-cada-minuto-infografia-infographic-socialmedia/

Blanco, D., (2016*). Para 2020 se triplicará información en centros de datos: Cisco*. Recuperado de https://www.elfinanciero.com.mx/tech/para-2020-habra-15-zettabytes-de-trafico-de-informacion-cisco/

Girones, J., (2013). *Data mining*. Recuperado de https://docplayer.es/146733360-Data-mining-jordi-girones-roig-pid_.html

Contacto

Carlos Alberto González Martínez

Jefe de departamento de correlaciones, cruces y alertas (C5i)

gmcmmiv@hotmail.com