

CIFF Trustees:



# Business Intelligence & Data Mining

Profesor:  
**Pedro Pasquau**

Octubre 2015

MASTER EN BUSINESS ANALYTICS & BIG DATA  
2015– 2016

# Objetivos

- **Sesión 1A: Introducción al Business Intelligence**
  - ☐ Business Intelligence
  - ☐ BI, BA y Big Data: Contexto
  - ☐ Pentaho BA
- **Sesión 1B ETL y Adquisición de Datos**
  - ☐ ETL
  - ☐ Pentaho Data Integration
  - ☐ Data Integration y Data Mining: Realimentación
- **Sesión 2A: Introducción Data Mining – Machine Learning**
  - ☐ Algoritmos
  - ☐ Ejemplos
- **Sesión 2B: Pentaho Data Scientist (PDS)**
  - ☐ PDS (Weka)
  - ☐ DW→ETL→PDS
  - ☐ PDS → ETL→DW

## A.- Data Mining

- Qué es Data Mining.
- Procesos de Data Mining
- Algoritmos de Data Mining
- Herramienta de Data Mining: Introducción a WEKA



## Objetivos

- Reconocer la problemática del análisis de grandes volúmenes de datos y de los beneficios de su uso sistemático para la obtención de modelos y patrones predictivos o descriptivos.
- Conocer las fases del Descubrimiento de Conocimiento de Bases de Datos y la importancia de las mismas en el éxito del proceso (en especial las de limpieza y selección de datos).
- Introducir las distintas técnicas de aprendizaje automático y estadísticas utilizadas en minería de datos, su potencial, su coste computacional y sus limitaciones de representación y de inteligibilidad.

## Objetivos

- Elegir, para un problema concreto, qué técnicas de minería de datos son más apropiadas.
- Generar los modelos y patrones elegidos utilizando una herramienta o paquete de minería de datos.
- Evaluar la calidad de un modelo, utilizando técnicas sencillas de evaluación (validación cruzada).
- Utilizar métodos de combinación de técnicas y de reiteración (p.ej. boosting).

## A.- Data Mining

- Qué es Data Mining.
- Proceso de Data Mining
- Algoritmos de Data Mining
- Herramienta de Data Mining: Introducción a WEKA
- Ejemplos sencillos sobre distintos algoritmos y análisis de los resultados obtenidos



## Causas del Origen del DM

### Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- El aumento del volumen y variedad de información que se encuentra almacenada en bases de datos digitales ha crecido espectacularmente en los últimos años. Nuevas tecnologías como BigData fomentan su crecimiento.
- Gran parte de esta **información es histórica**, es decir, representa transacciones o situaciones que se han producido.
- Aparte de su función de “memoria de la organización”, la información histórica es útil **para predecir la información futura**.



## Causas del Origen del DM

- La mayoría de *decisiones* de empresas, organizaciones e instituciones se basan también en información de experiencias pasadas extraídas de fuentes muy diversas.
- las **decisiones colectivas** suelen tener consecuencias mucho más graves, especialmente económicas, y, recientemente, se deben basar en **volúmenes de datos que desbordan la capacidad de la mente humana**.

## Causas del Origen del DM

- Tamaño de datos poco habitual para algoritmos clásicos:
  - número de registros (ejemplos) muy largo ( $10^8$ - $10^{12}$  bytes).
  - datos altamente dimensionales ( $n^\circ$  de columnas/atributos).
- El usuario final no es un experto en aprendizaje automático ni en estadística.
- El usuario no puede perder más tiempo analizando los datos:
  - industria: ventajas competitivas, decisiones más efectivas.
  - ciencia: datos nunca analizados, bancos no cruzados, etc.
  - personal: “information overload”...

## Causas del Origen del DM

- *Los procesos-sistemas clásicos de estadística son difíciles de aplicar-usar y no escalan al número de datos típicos en bases de datos actuales.*
- *Crítico en ecosistemas BigData o Sistemas Distribuidos*

## DataMinig?

Surge KDD!!!

- “Descubrimiento de Conocimiento implícito en Bases de Datos” (KDD, del inglés *Knowledge Discovery from Databases*).
- “*proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos*”. Fayyad (1996)
- Diferencia clara con métodos estadísticos: la estadística se utiliza exclusivamente para **validar** o parametrizar un *modelo sugerido y preexistente*, no para **generarlo**.
- Diferencia sutil “Análisis Inteligente de Datos” (IDA, del inglés Intelligent Data Analysis) que correspondía con el uso de técnicas de inteligencia artificial en el análisis de los datos.

## Qué es DataMining y qué NO es

- KDD nace como Suite completa o Concepto Global y se nutre de diferentes disciplinas:
  - estadística.
  - sistemas de información / bases de datos.
  - aprendizaje automático / IA.
  - visualización de datos.
  - computación paralela / distribuida.
  - interfaces de lenguaje natural a bases de datos.

## Qué es DataMining y qué NO es

- La minería o prospección de datos (DM) no es más que una **fase** del KDD:
  - Fase que integra los métodos de aprendizaje y estadísticos para obtener hipótesis de patrones y modelos.
- Al ser la fase de generación de hipótesis, vulgarmente se asimila KDD con DM.
- Además, las connotaciones \*\* del término “minería de datos” han hecho que éste se use como identificador del área.

## Qué es DataMining y qué NO es

La minería de datos **no** es una extensión de los sistemas de informes inteligentes o sistemas **OLAP** (*On-Line Analytical Processing*).

**La minería de datos aspira a más**

Otras herramientas, p.ej. consultas sofisticadas o análisis estadístico, pueden responder a preguntas como:

“¿Han subido las ventas del producto X en junio?”

“¿Las ventas del producto X bajan cuando promocionamos el producto Y?”

Pero sólo con técnicas de minería de datos podremos responder a preguntas del estilo:

“¿Qué factores influyen en las ventas del producto X?”

“¿Cuál será el producto más vendido si abrimos una delegación en Portugal?”

## DataMining

- Visión con las herramientas tradicionales:
  - El analista empieza con una pregunta, una suposición o simplemente una intuición y explora los datos y construye un modelo. **El analista propone el modelo.**  
Realiza preguntas
- Visión con la minería de datos:
  - Aunque el analista no pierde la posibilidad de proponer modelos, **el sistema encuentra y sugiere modelos.**



## DataMining

### Ventajas:

- Generar un modelo requiere menos esfuerzo manual y permite evaluar cantidades ingentes de datos.
- Se pueden evaluar muchos modelos generados automáticamente, y esto aumenta la probabilidad de encontrar un buen modelo.
- El analista necesita menos formación sobre construcción de modelos y menos experiencia.

## Áreas de Aplicación

Más importante industrialmente

Áreas de Aplicación:

- **Business Intelligence** -Toma de Decisiones (banca-finanzas-seguros, márketing, políticas sanitarias/demográficas, ...)
- **Procesos Industriales** (componentes químicos, compuestos, mezclas, esmaltes, procesos, etc.)
- **Investigación Científica** (medicina, astronomía, meteorología, psicología, ...). Aquí la eficiencia no es tan importante.

## Áreas de Aplicación

Áreas de Aplicación:

- Soporte al Diseño de Bases de Datos.
- *Reverse Engineering* (dados una base de datos, desnormalizarla para que luego el sistema la normalice).
- Mejora de Calidad de Datos.
- Mejora de Consultas (si se descubren dependencias funcionales nuevas u otras condiciones evitables).

## Ejemplos de Aplicación

### KDD para toma de decisiones

- Comercio/Marketing:**
- Identificar patrones de compra de los clientes.
  - Buscar asociaciones entre clientes y características demográficas.
  - Predecir respuesta a campañas de *mailing*.
  - Análisis de cestas de la compra.

- Banca:**
- Detectar patrones de uso fraudulento de tarjetas de crédito.
  - Identificar clientes leales.
  - Predecir clientes con probabilidad de cambiar su afiliación.
  - Determinar gasto en tarjeta de crédito por grupos.
  - Encontrar correlaciones entre indicadores financieros.
  - Identificar reglas de mercado de valores a partir de históricos.

## Ejemplos de Aplicación

### **KDD para toma de decisiones**

#### Seguros y Salud Privada:

- Análisis de procedimientos médicos solicitados conjuntamente.
- Predecir qué clientes compran nuevas pólizas.
- Identificar patrones de comportamiento para clientes con riesgo.
- Identificar comportamiento fraudulento.

#### Transportes: GEO

- Determinar la planificación de la distribución entre tiendas.
- Analizar patrones de carga.

## Ejemplos de Aplicación

### Medicina: Tratamiento Digital de Imágenes Médicas

- Identificación de terapias médicas satisfactorias para diferentes enfermedades.
- Asociación de síntomas y clasificación diferencial de patologías.
- Estudio de factores (genéticos, precedentes, hábitos, alimenticios, etc.) de riesgo/salud en distintas patologías.
- Segmentación de pacientes para una atención más inteligente según su grupo.
- Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
- Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, etc.

## Ejemplos de Aplicación

### **Procesos Industriales**

- Extracción de modelos sobre comportamiento de compuestos.
- Detección de piezas con trabas.
- Predicción de fallos
- Modelos de calidad.
- Estimación de composiciones óptimas en mezclas.
- Extracción de modelos de coste.
- Extracción de modelos de producción.
- Simulación costes/beneficios según niveles de calidad