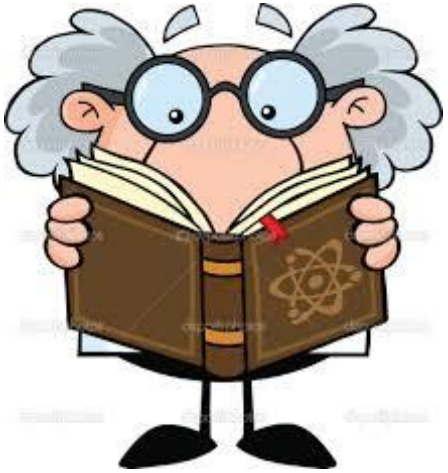


Introducción a la Minería de Datos

Dra. Irene Olaya Ayaquica Martínez
Facultad de Ciencias de la Computación, BUAP
irene.ayaquica@correo.buap.mx

Objetivos

- Antecedentes
- ¿Qué es minería de datos?
- Historia
- Características y objetivos
- Ciclo virtuoso de la minería de datos
- Usos de la minería de datos
- Tendencias
- Herramientas de software
- ¿Por qué de la minería de datos?

Antecedentes

- En las últimas décadas, los avances en el poder y la velocidad de procesamiento nos han permitido pasar de las prácticas manuales, tediosas y que toman mucho tiempo al análisis de datos rápido, fácil y automatizado.

- Cuanto más complejos son los conjuntos de datos recopilados, mayor es el potencial que hay para descubrir insights relevantes.

Antecedentes

- Frecuentemente hay información "**oculta**" en los datos, que en realidad no es evidente.
- A los analistas humanos les puede tomar semanas **descubrir** información que sea útil.
- Mucha de esta información no es **analizada** del todo.
- La extracción de información se vuelve un tema relevante.

¿Qué es minería de datos?

- La minería de datos consiste en la extracción no trivial de información, que reside de manera implícita en los datos. Dicha información, previamente desconocida, podrá resultar útil para algún proceso.

- La minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

¿Qué es minería de datos?

- La minería de datos es una colección de técnicas para el descubrimiento automatizado eficiente de patrones previamente desconocidos, válidos, novedosos, útiles y comprensibles en grandes bases de datos.

- Los patrones deben ser accionables para que puedan usarse en la toma de decisiones de una empresa.

¿Qué es minería de datos?

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

- La minería de datos estudia métodos y algoritmos que permiten la extracción automática de información sintetizada que permite caracterizar las relaciones escondidas.

Historia

- 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.

Historia

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

- Las listas de discusión sobre este tema las forman investigadores de más de ochenta países.

- Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

Características y objetivos

- Las empresas están principalmente interesadas en descubrir patrones pasados para predecir el comportamiento futuro.
- Un almacén de datos puede ser la memoria de una empresa. La minería de datos puede proporcionar inteligencia usando esa memoria.
- Asumimos que estamos tratando con grandes cantidades de datos, quizás Gigabytes o quizás Terabytes.

Características y objetivos

- Aunque la minería de datos es posible con una menor cantidad de datos, cuanto mayor es la información, mayor es la confianza en cualquier patrón desconocido que se descubra.

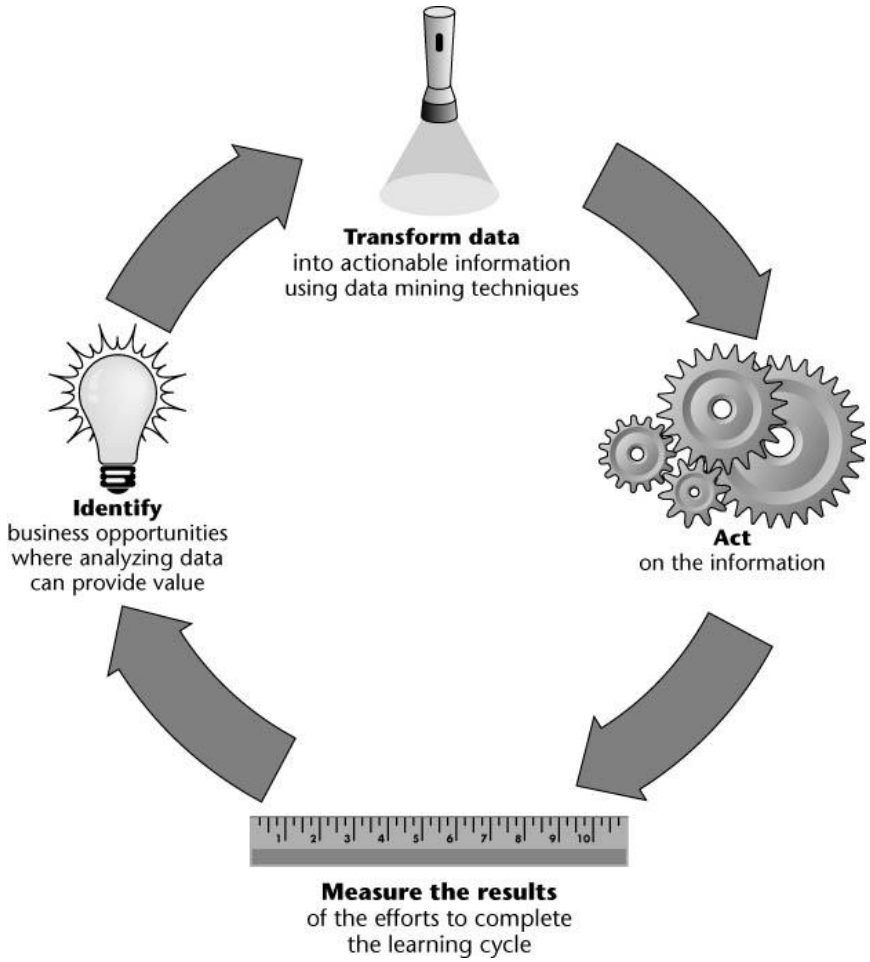
- No existe un solo enfoque para minería de datos sino un conjunto de técnicas que se pueden utilizar de manera independiente o en combinación.

Características y objetivos

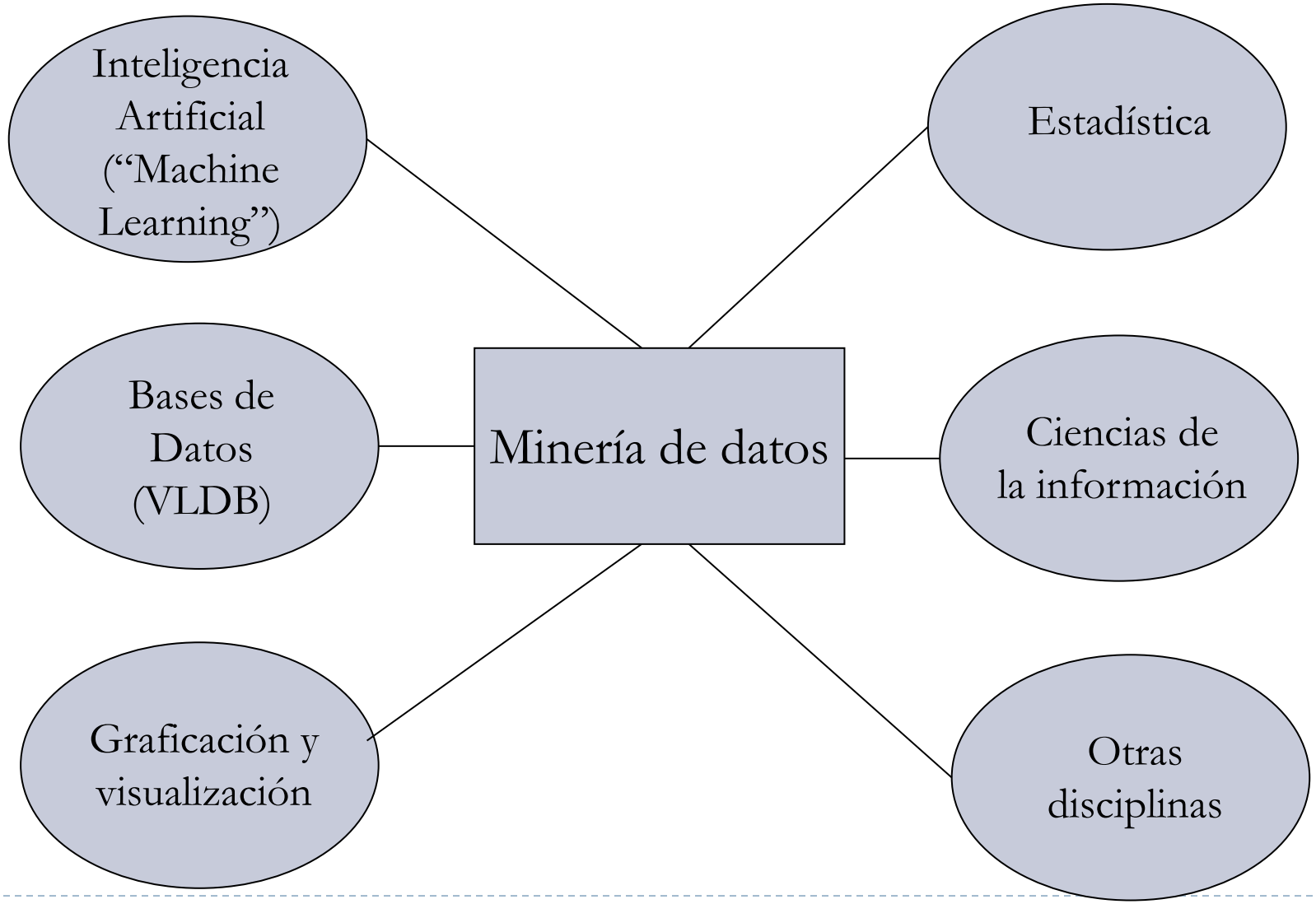
- Permitir a una organización MEJORAR _____ a través de un mejor CONOCIMIENTO de _____

- Mejorar la ventaja competitiva

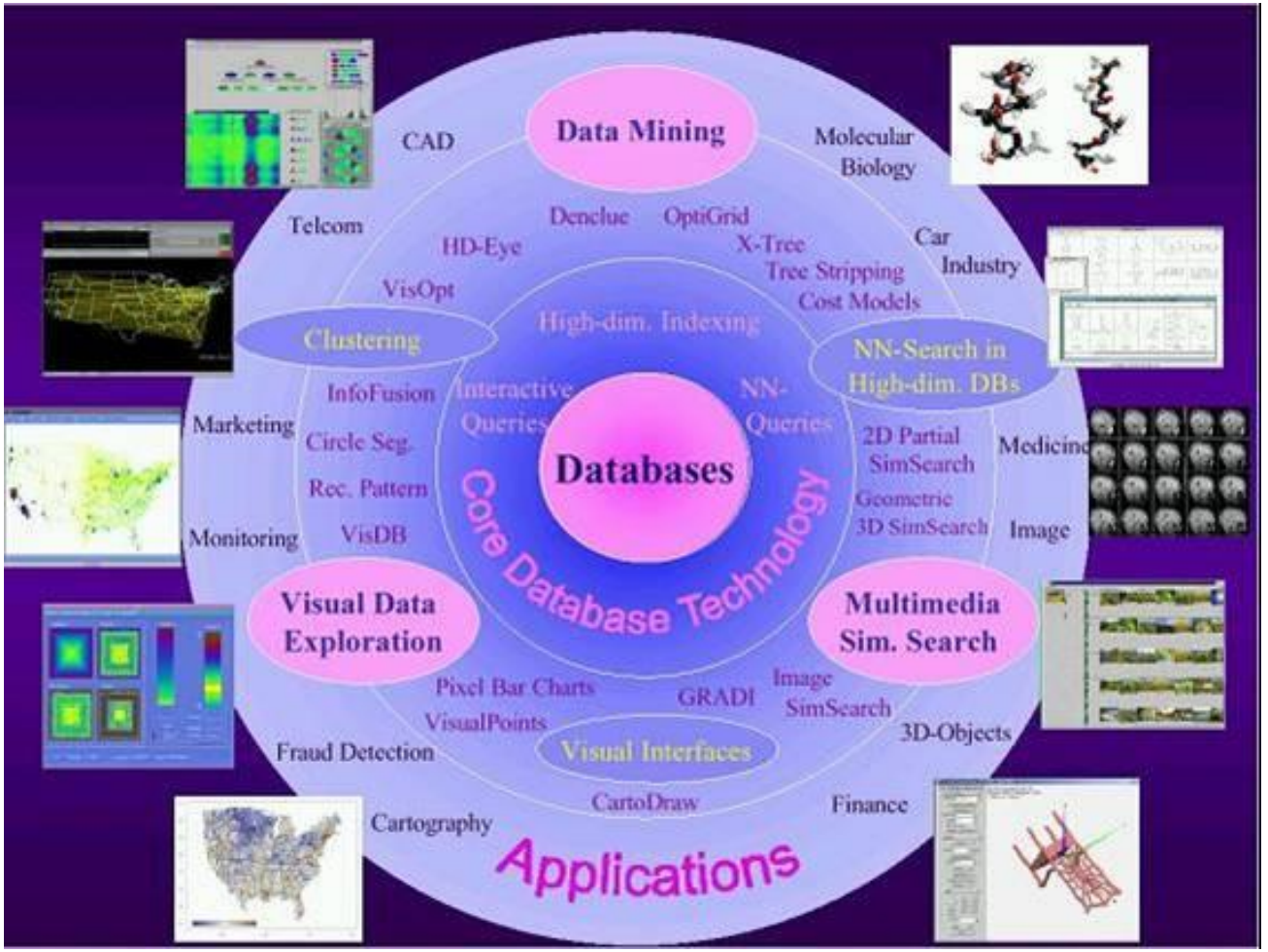
Ciclo virtuoso de la minería de datos



La minería de datos es un campo multidisciplinario



La minería de datos es un subconjunto de la inteligencia de negocios



Tipos de aplicaciones de la minería de datos

- Las aplicaciones o problemas de minería de datos pueden clasificarse en las siguientes categorías:
  - Clasificación
  - Estimación
  - Pronóstico
  - Asociación
  - Agrupación o segmentación

Clasificación

- Examinar las características de un nuevo objeto y asignarle una clase o categoría de acuerdo a un conjunto de tales objetos previamente definido.

- Ejemplos:
  - Clasificar aplicaciones a crédito como bajo, medio y alto riesgo
  - Detectar reclamos fraudulentos de seguros

Estimación

- Relacionado con clasificación
  - Mientras clasificación asigna un valor discreto, estimación produce un valor continuo.

- Ejemplos:
  - Estimar el precio de una vivienda
  - Estimar el ingreso total de una familia

Pronóstico

- Predecir un valor futuro con base a valores pasados.

- Ejemplos:
  - Predecir cuánto efectivo requerirá un cajero automático en un fin de semana

Asociación

- Determinar cosas u objetos que van juntos.

- Ejemplo:
  - Determinar que productos se adquieren conjuntamente en un supermercado

Agrupación o segmentación

- Dividir una población en un número de grupos más homogéneos.
- No depende de clases pre-definidas a diferencia de clasificación.
- Ejemplo:
  - Dividir la base de clientes de acuerdo con los hábitos de consumo

Usos de la minería de datos

- Negocios
  - Administración empresarial
- Patrones de fuga
- Recursos humanos
  - Dirección estratégica
- Comportamiento en internet
  - E-commerce

Usos de la minería de datos

- Terrorismo
- Juegos
- Ciencia e ingeniería
  - Genética
  - Ingeniería eléctrica
  - Análisis de gases

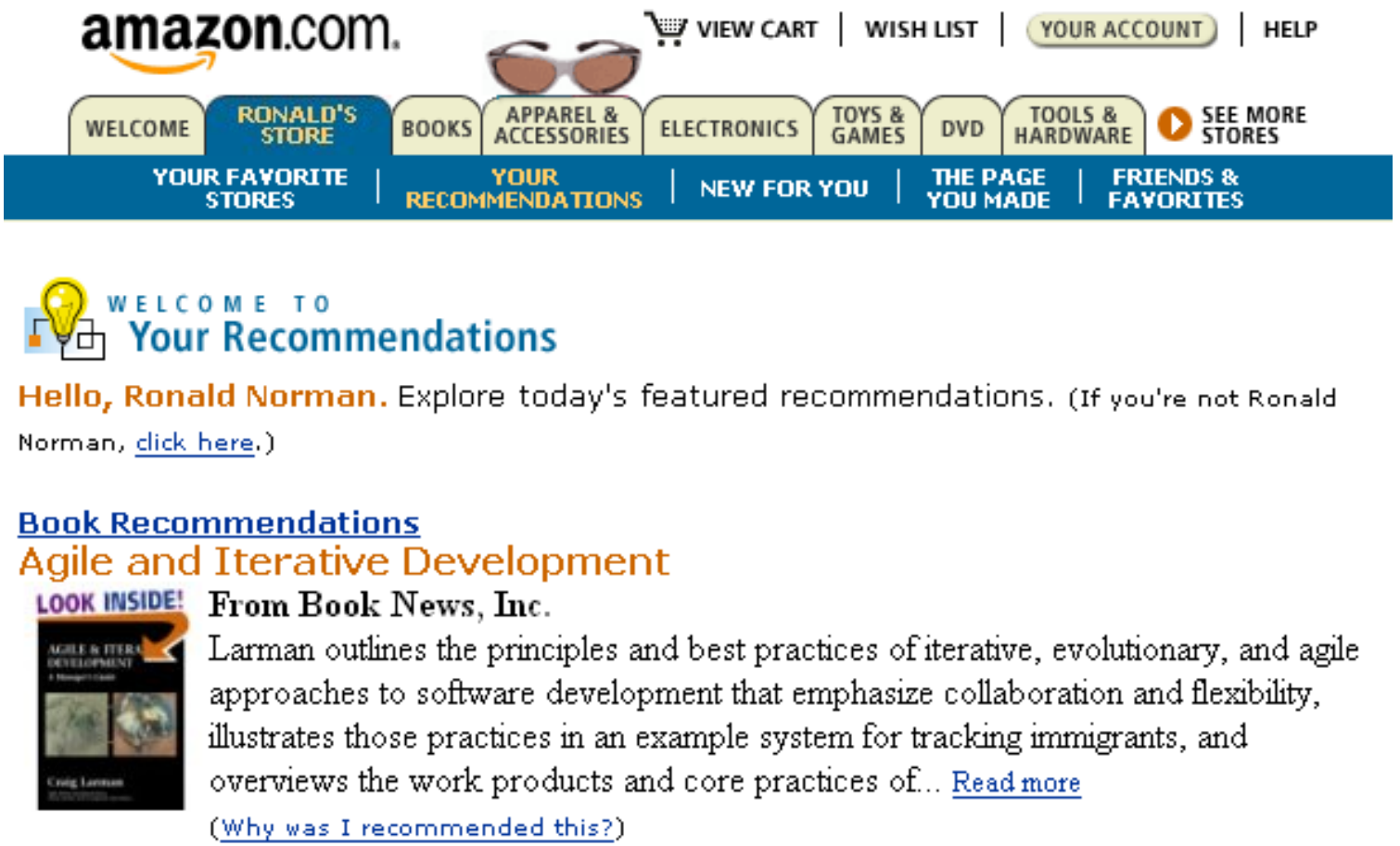
Usos de la minería de datos

- Comunicaciones
- Seguros
- Educación
- Manufactura
- Bancos

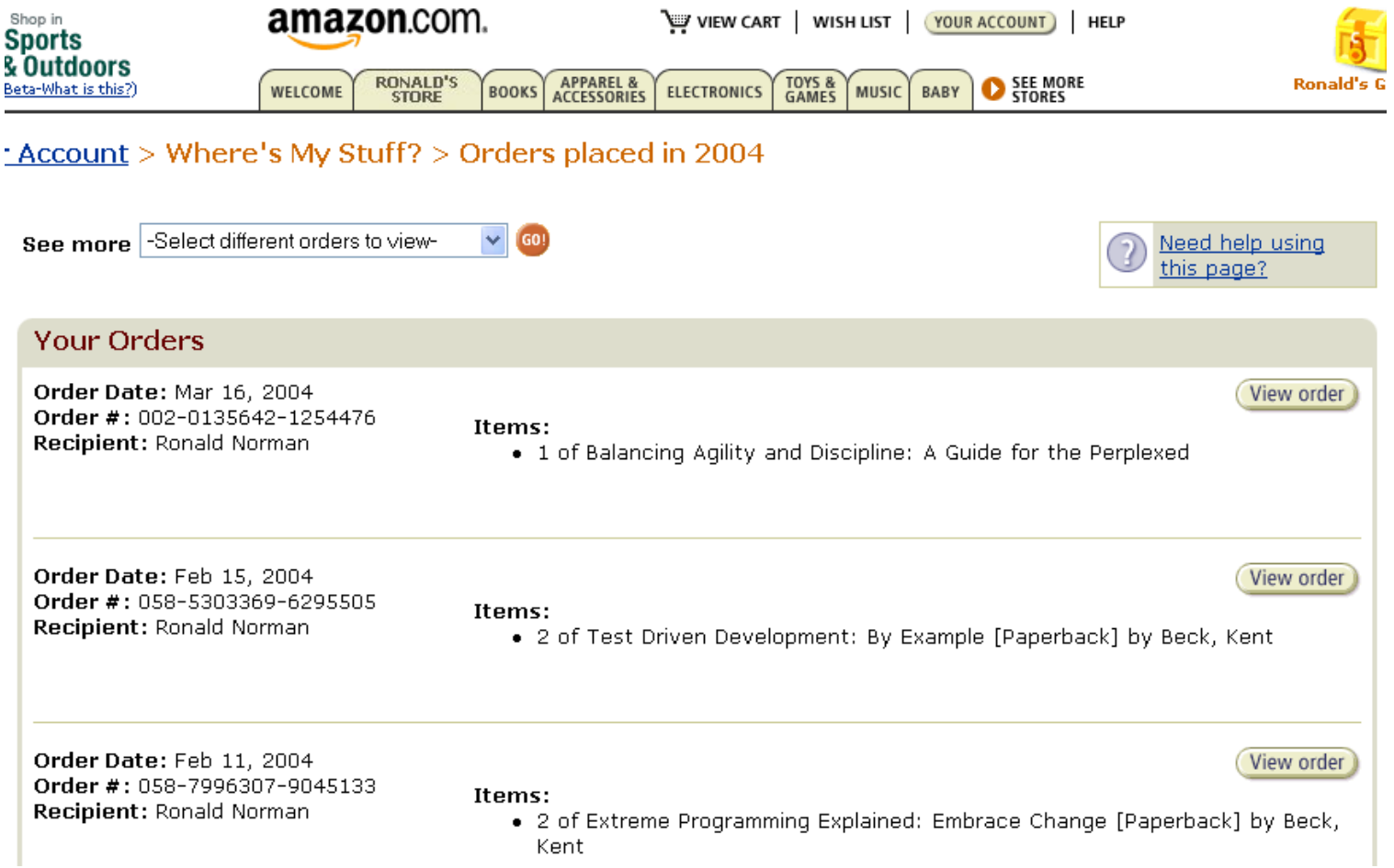
Ejemplos

- amazon.com utiliza asociaciones.
- Las recomendaciones a los clientes se basan en compras pasadas y en lo que otros clientes están comprando.

Recomendaciones



Disponibilidad de datos de transacciones



Ejemplos

- Una tienda en EE. UU. "Just for Feet" tenía alrededor de 200 tiendas, cada una con hasta 6000 estilos de zapatos, cada estilo en varios tamaños.

- La minería de datos se utilizaba para encontrar los zapatos correctos para almacenar en la tienda correcta.

Tendencias

- Las tendencias actuales tienen que ver con las transformaciones que ha sufrido la minería de datos. Las principales son:

  - La importancia que han cobrado los datos no estructurados (texto, páginas de Internet, etc.).
  - La necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.
  - La exigencia de que los procesos funcionen prácticamente en línea.
  - Los tiempos de respuesta.

Herramientas de software

- Dynamic Data Web
- KXEN
- KNIME
- Orange
- RapidMiner
- R-project
- SPSS
- SAS Enterprise Miner
- STATISTICA Data Miner
- **Weka**

¿Por qué de la minería de datos?

- Datos se encuentran disponibles
- Poder computacional es cada vez menos costoso
- Las presiones competitivas son enormes
- Software para minería de datos se encuentra disponible

¿Por qué de la minería de datos?

- Crecimiento en la generación y almacenamiento de datos corporativos - explosión de información
- El volumen de datos producidos se duplica cada dos años.
- Los datos no estructurados por sí solos conforman el 90% del universo digital. Pero más información no significa necesariamente más conocimientos.

¿Por qué de la minería de datos?

- Evolución de la tecnología: almacenamiento mucho más barato, recolección de datos más fácil, mejor administración de la base de datos para el análisis y la comprensión de los datos.

- Necesidad de una toma de decisiones sofisticada: los sistemas de bases de datos actuales son sistemas de procesamiento de transacciones en línea (OLTP). Los datos OLTP son difíciles de usar para tales aplicaciones.