UNIVERSIDAD AUSTRAL | INGENIERÍA

Introducción a Data Mining

Dr. Juan M. Ale

Mg. Gaston Pezzuchi, MSc.

Que es data mining?

- Data mining es el proceso de descubrir patrones y tendencias útiles en grandes conjuntos de datos… (Larose, 2015)
  - De acuerdo con el Gartner Group
    - La mayoría de las compañías americanas con mas de 1000 empleados tienen por lo menos 200TB de datos, y los mismos crecen a un 40 % anual.
    - Las compañías de *retail* pueden incrementar sus márgenes operativos en mas de un 60 %...
    - El sector de salud puede reducir fuertemente sus costos, mejorando a la vez su eficiencia y calidad…
  - Elecciones Presidenciales EEUU 2012 (source: MIT Technology Review - https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/)
    - First identified likely Obama voters using a data mining model, and then made sure that these voters actually got to the polls
    - used a separate data mining model to predict the polling outcomes county-by-county
    - Hamilton, Ohio: the model predicted 56.4% for Obama; actual result was 56.6%, so that the prediction was off by only 0.02%

Wanted: Data Miners

"We are drowning in information but starved for knowledge."
Megatrends, John Naisbitt

- We are inundated with data in most fields, but…
- There are not trained human analysts available who are skilled to convert the data into knowledge
- According to McKinsey Report
  - "There will be a shortage of talent…"
  - "…particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data."
  - Demand for talent to exceed supply "…by 140,000 to 190,000 positions"
  - "… we project a need for 1.5 million additional managers and analysts in the United States"

- Factors
  - Explosive growth in data collection, as in supermarket scanners
  - Storing the data in data warehouses
  - Increased access to data from web navigation an intranets
  - Competitive pressure to increase market share in globalized economy
  - Growth of computing power and storage capacity

La necesidad de la Dirección Humana en las tareas de Data Mining

- Algunas definiciones iniciales sobre la minería de datos, la definían como un proceso "automático"
- "*…this has misled many people into believing data mining is product that can be bought rather than a discipline that must be mastered.*" (Berry, Linoff)
- La Automatización **NO ES SUSTITUTO** para la intervención humana.
- Es muy fácil "hacer" minería de datos en forma **incorrecta**.
- Es un **requisito** entender las estructuras y modelos estadísticos y matemáticos que subyacen en las implementaciones de software.
- Es **necesaria** la intervención humana activa e involucrada en cada una de las fases del proceso de minería de datos.
- Las tareas de minería de datos deben integrarse en un proceso humano de resolución de problemas.

Que es data mining?

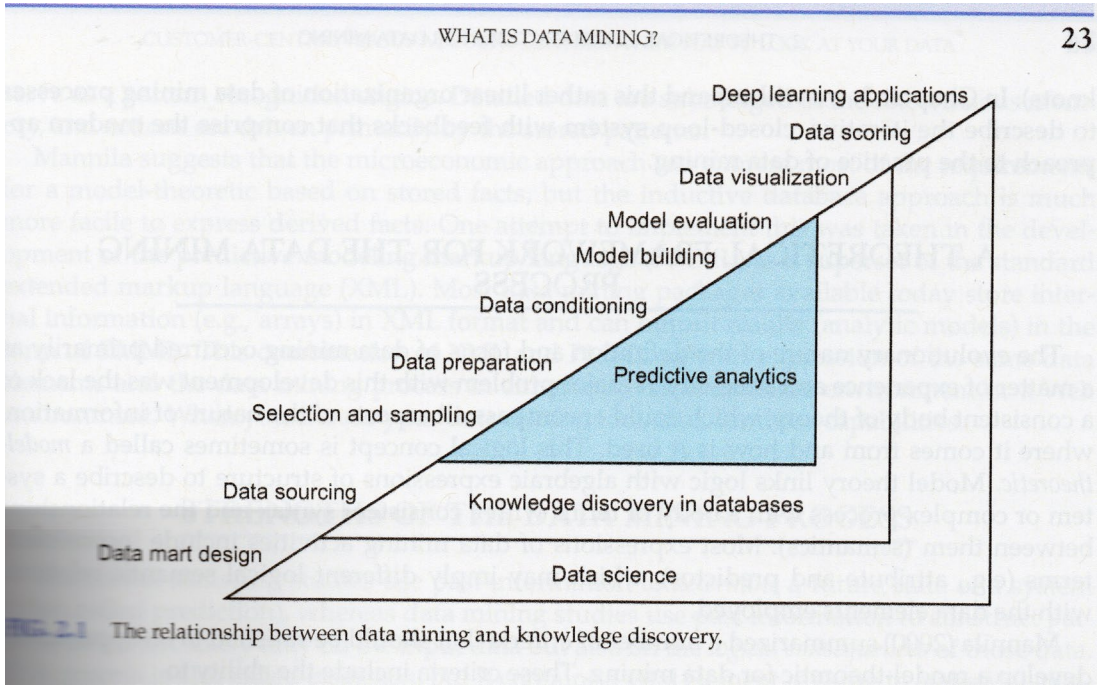
- La minería de datos puede ser definida de diferentes formas que difieren en su foco. Una de las primeras definiciones fue:
  - The *non-trivial* extraction of *implicit,* previously unknown, and potentially useful information from data. ***(Frawley et at., 1991)***

- En la medida en que la minería de datos se desarrolló como una actividad profesional, fue necesario distinguirla de otras actividades como el modelado estadístico o el descubrimiento de conocimiento. Nisbet *et al.* 2018, proponen:

  - *Statistical modeling:* The use of parametric statistical algorithms to group or predict an outcome or event, based on predictor variables.
  - *Data mining:*The use of machine-learning algorithms to find faint patterns of relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benefit in some form (diagnosis, profit, detection, etc.).
  - *Knowledge Discovery:* The entire process of data Access, data exploration, data preparation, modeling, model deployment and model monitoring. This broad process includes data mining activities.
  - *Data Science:* The extensión of knowledge Discovery into data architecture of analytic data marts on one hand and complex image, speech, and textual analysis on the other hand with highly evolved machine-learning algorithms.

Que es data mining?

- En la medida en que la practica de la minería de datos se fue desarrollando, la definición cambió a aspectos específicos de la información y de las fuentes. En 1996, Fayyad *et al.* propusieron:
  - Knowledge Discovery in databases is the non-trivial process of identifying valid, novel, potential useful, and ultimately understandable patterns in data.



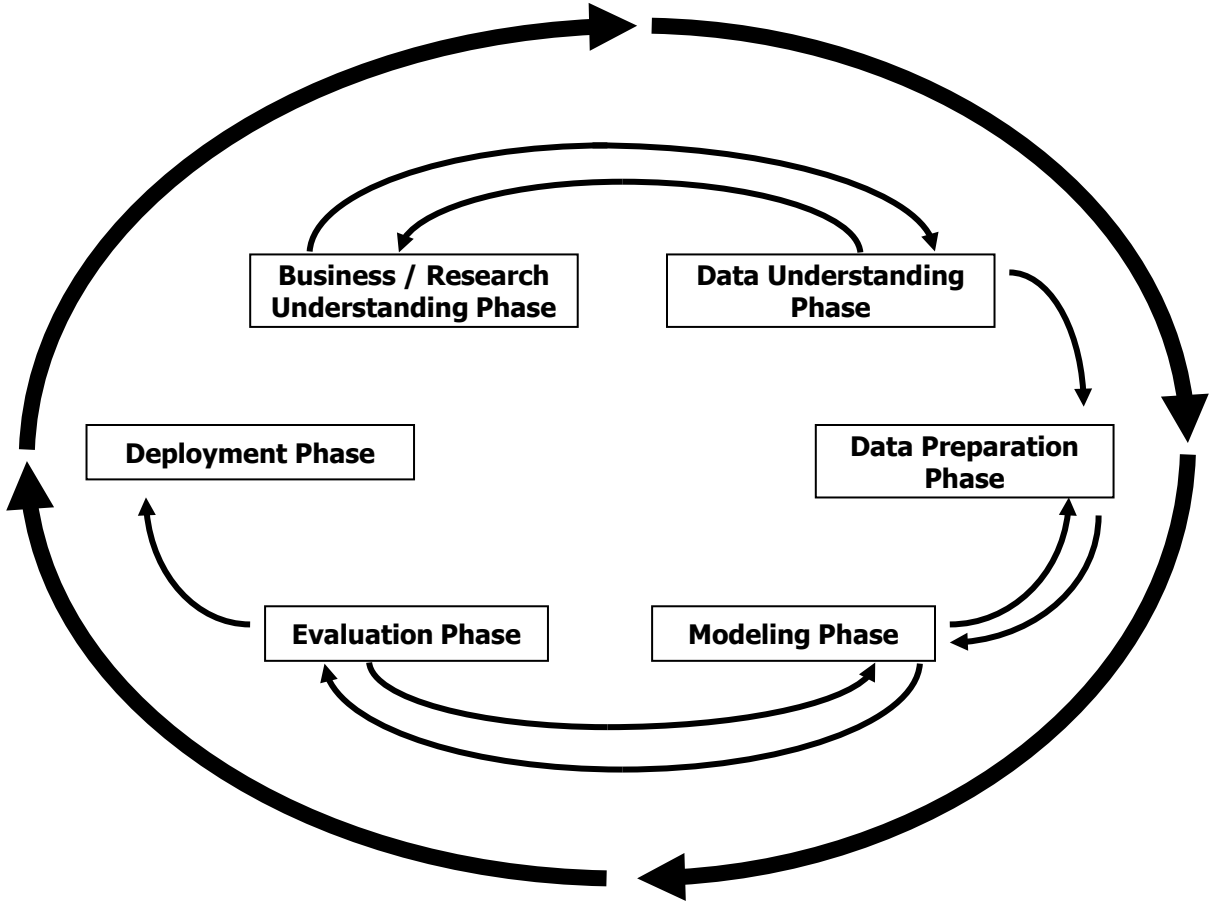
The relationship between data mining and knowledge discovery.

Cross Industry Standard Process: CRISP-DM

- El Cross-Industry Standard Process for Data Mining (CRISP-DM) fue desarrollado en 1996
  - Considera a la minería de datos dentro de una estrategia general de resolución de problemas para una unidad de investigación o para un negocio.
  - En su desarrollo contribuyeron entre otros: DaimlerChrysler, SPSS, y NCR
  - Se lo considera un proceso neutro en las dimensiones
    - Industria
    - Herramienta
    - Aplicación
  - Es un proceso No propietario y disponible en forma abierta.
  - Los proyectos de minería de datos siguen un ciclo iterativo, adaptativo de 6 fases.
  - Las fases secuenciales son **adaptativas.**

Cross Industry Standard Process: CRISP-DM

- El círculo exterior muestra el proceso iterativo.
- Solo se muestran las dependencias mas significativas entre las fases.
- La fase siguiente depende de los resultados de la fase precedente.
- Es posible Volver a una fase anterior antes de avanzar a la siguiente.



CRISP-DM Lifecycle

Cross Industry Standard Process: CRISP-DM

- (1) Fase de Comprensión del Negocio/Investigación
  - Define los requerimientos y objetivos del proyecto
  - Traduce los objetivos en una definición del problema de minería de datos.
  - Prepara una estrategia preliminar para cumplir con esos objetivos
- (2) Fase de Comprensión de Datos
  - Reunir los datos
  - Realizar Análisis Exploratorio de Datos (AED) (Exploratory data analysis (EDA))
  - Evaluar la calidad de los datos
  - Opcionalmente, seleccionar los subconjuntos interesantes
- (3) Fase de Preparación de Datos
  - Prepara los datos para el modelamiento en las fases siguientes
  - Selecciona casos y variables apropiadas para análisis
  - Limpia y prepara los datos de manera que son adecuados para las herramientas de modelado
  - Si es necesario, realiza transformaciones de ciertas variables.

Cross Industry Standard Process: CRISP-DM

- (4) Fase de Modelado
  - Selecciona y aplica una o varias técnicas de modelización
  - Calibra el modelo para optimizar los resultados
  - De ser necesario, puede realizar tareas adicionales de preparación de datos para apoyat una técnica de modelización específica.
- (5) Fase de Evaluación
  - Evalúa uno o mas modelos en relación con su efectividad
  - Determina si se han cumplido o no los objetivos
  - Establece si alguna faceta importante del problema no ha sido suficientemente tenida en cuenta
  - Toma decisiones en relación con los resultados de la mineria de datos previo a su despliegue en el campo

Cross Industry Standard Process: CRISP-DM

- (6) Fase de Despliegue
  - Utiliza los modelos que han sido creados
  - Un ejemplo simple de despliegue: Genera un reporte
  - Un ejemplo mas complejo: implementa un esfuerzo de minería de datos en paralelo en otro departamento
  - En los negocios, usualmente se utiliza el modelo para decidir las estrategias de respuesta.

Falacias de la Mineria de Datos

- Four Fallacies of Data Mining (Louie, Nautilus Systems, Inc.)

	Falacias	Realidad
1	• Es posible desplegar sin planificación un conjunto de herramientas en repositorios de datos • Encuentra soluciones a todos los problemas de negocios	• No hay herramientas automáticas de minería de datos que resuelvan problemas. • Si hay diferentes procesos de minería de datos. • Se integra a los objetivos generales de negocios.
2	• El proceso de minería de datos es autónomo • Requiere poca supervisión	• Requiere una intervención humana significativa en cada fase del proceso. • Luego de que el modelo se despliega, nuevos modelos requieren actualizaciones. • Los analistas monitorean y evalúan continuamente
3	• La minería de datos se paga sola rápidamente	• Las tasas de retorno varian.Return rates vary • Dependen del punto de inicio, el personal, los costos de preparación de datos, etc.
4	• El software de minería de datos es fácil de usar.	• La facilidad de uso, varia entre proyectos. • Los analistas deben combinar el conocimiento del tema con conocimiento especifico del dominio del problema.

Falacias de la Mineria de Datos

- Otras Falacias (Larose)

	Falacia	Realidad
5	• La minería de datos identifica la cauda de los problemas de negocios.	• El proceso de descubrimiento de conocimiento es el que encuentra patrones de comportamiento • Los seres humanos interpretan los resultados e identifican las causas
6	• La minería de datos automáticamente limpia los datos en las bases de datos.	• La minería de datos muchas veces utiliza datos de sistemas heredados (*legacy*) • Es posible que esos datos no hayan sido examinados en años • Las organizaciones suelen iniciar tareas de minería de datos confrontándose con gigantescas tareas de preprocesamiento de datos.
7	• La minería de datos siempre provee resultados positivos.	• No hay garantías de resultados positivos. • Pero si la minería de datos se realiza adecuadamente, se obtendrán siempre resultados “accionables” y muchas veces redituables.