Presentación

- “Tópicos de minería de datos”
  - Materia optativa de LCC
  - Materia de Doctorado en Informática

- Docente: Pablo M. Granitto
- Horarios: Todo virtual!

Introducción

- Qué es la minería de datos?

- Es lo mismo que Machine Learning?

Introducción

- Data mining:

“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”

From: Principles of Data Mining. Hand, 2001

Introducción

- Data mining:

“Data mining is the extraction of implicit, previously unknown, and potentially useful information from data”

From: Data Mining. Witten, 2005

Introducción

- Pattern Recognition:

“The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories.”

From: Pattern Recognition and Machine Learning, Bishop, 2006

Introducción

- Son lo mismo?
  - Para algunos autores son lo mismo.
  - Hay diferencias:
    - PR se concentra en problemas de clasificación, en especial en Machine Vision y reconocimiento de escritura/voz
    - DM es más amplio, incorpora más problemas (regresión, reglas, etc.) y más etapas del proceso (pre y post procesado)

Introducción

- Data Science:

“Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.”

From: Wikipedia (really: doi:10.1145/2500499)

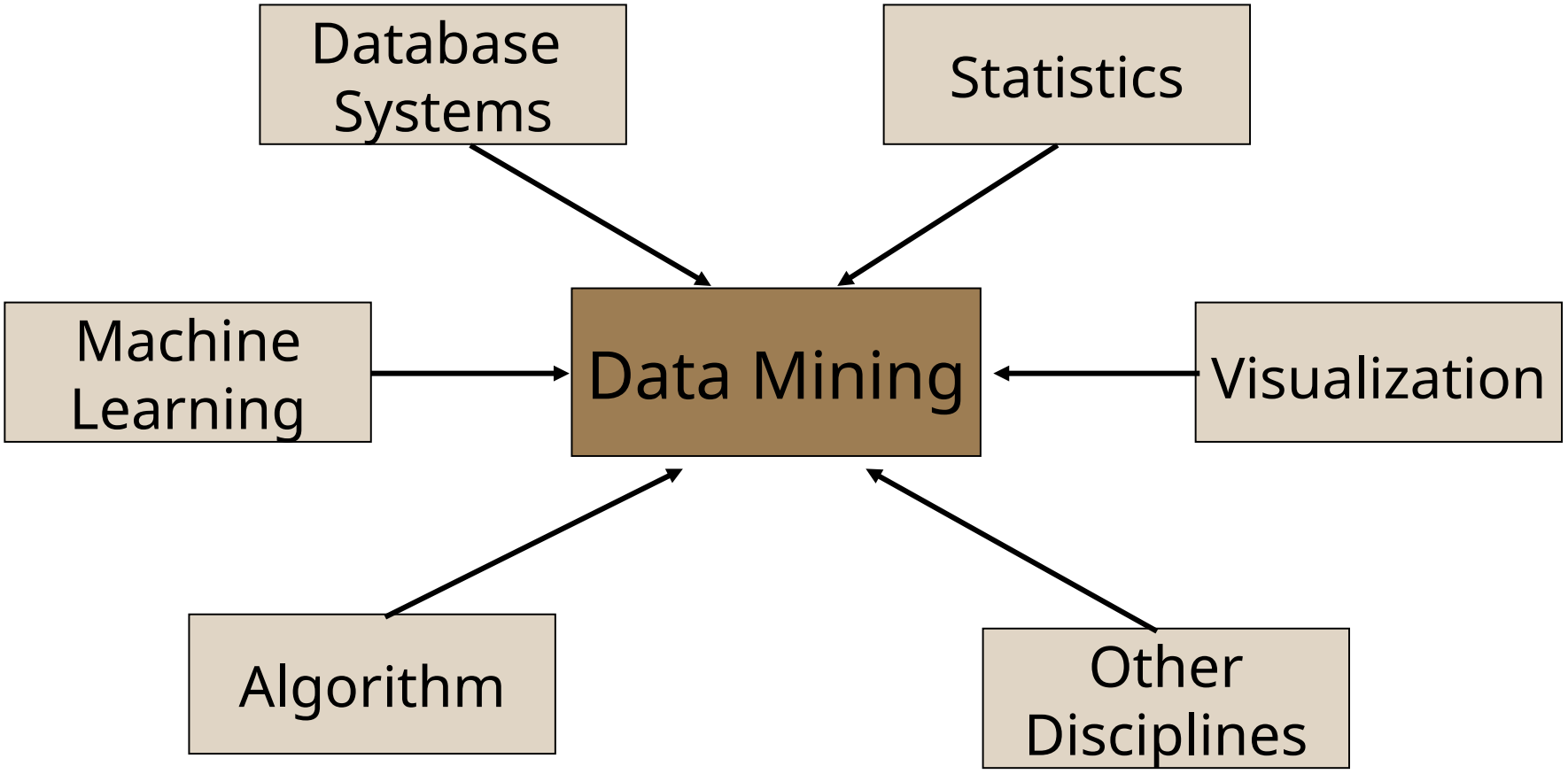
Introducción

- Son lo mismo?
  - Si, en general hay consenso en que *“Data Science”* es una forma *“cool”* de hablar de *“Data Mining”*

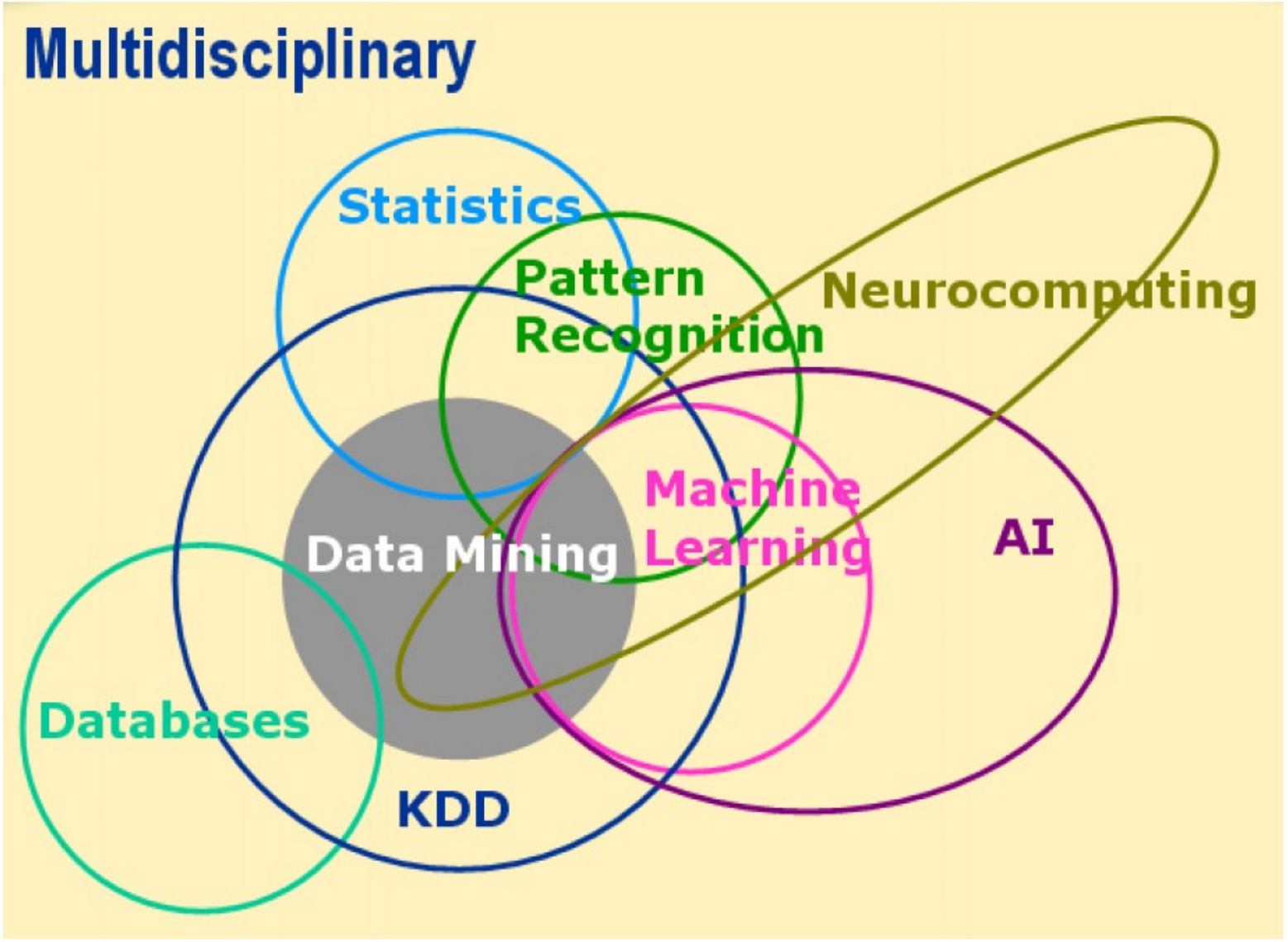
Otros nombres

- Knowledge discovery (mining) in databases (KDD)
- Knowledge extraction
- Data/pattern analysis
- Data archeology
- Data dredging
- Information harvesting
- Business intelligence
- etc.

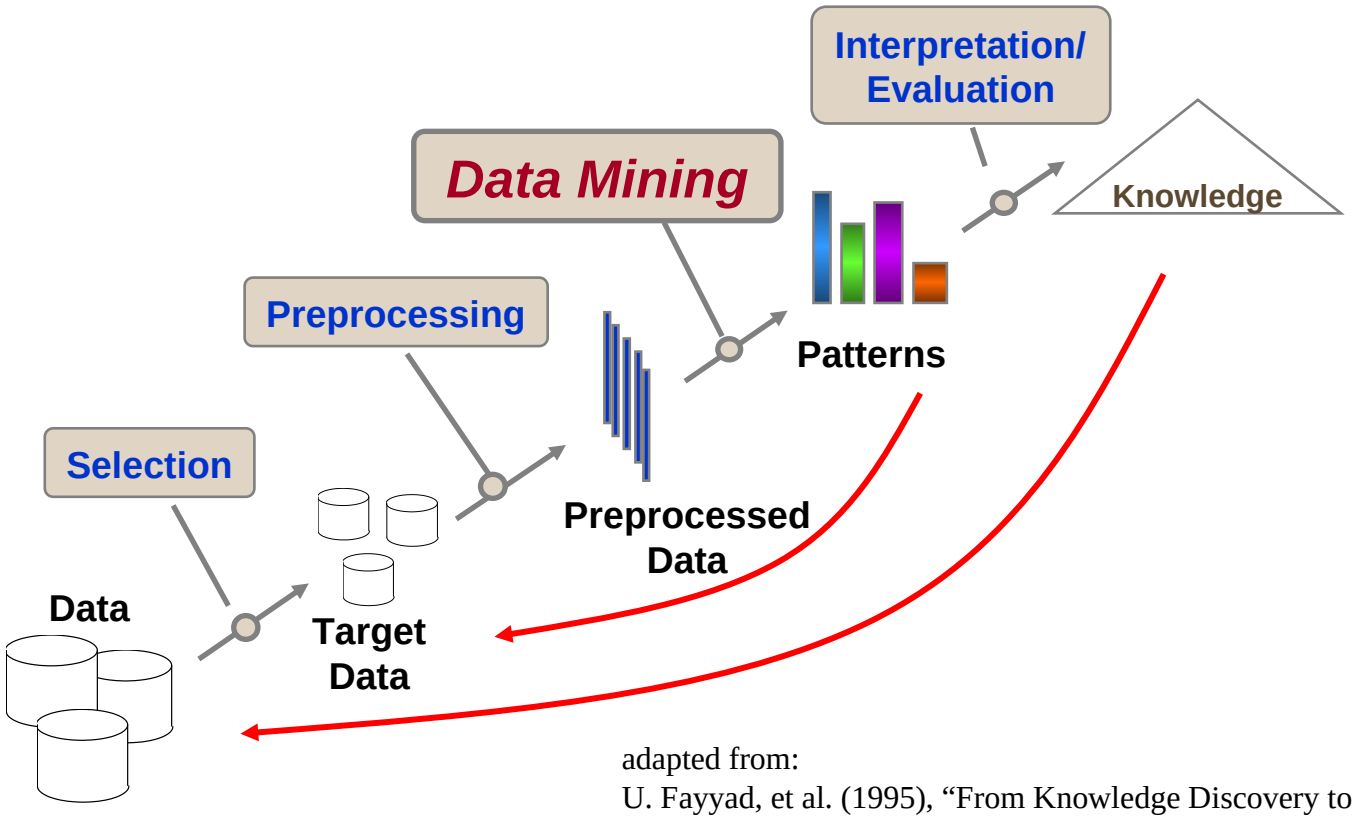
Confluencia de disciplinas



Confluencia de disciplinas



La visión “KDD”



adapted from:
U. Fayyad, et al. (1995), “From Knowledge Discovery to Data Mining: An Overview,” Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

En esta visión Data Mining se limita a la aplicación práctica de los métodos de Machine Learning sobre grandes bases de datos. No es nuestra visión!

Por qué DM?

- Estamos completamente sobrepasados por la cantidad de datos a nuestra disposición.
- Esa cantidad se incrementa en forma exponencial: sistemas automáticos de almacenamiento en comercios, industria y ciencia.
- Mientras los datos crecen, la proporción de los mismos que la gente puede entender decrece proporcionalmente (alarmante)

Por qué DM?

- Datos e información son dos conceptos distintos (información y conocimiento también).
- “Nos ahogamos en datos peros seguimos sedientos de información”
- Dentro de nuestros datos está oculta la información, pero no puede ser aprovechada si no se la “extrae” de los datos

Buscar patrones no es nuevo!

- El cerebro humano busca patrones en todos lados.
  - Desde el principio de la raza humana. Los cazadores buscaban patrones en el comportamiento de los animales, los agricultores en el crecimiento de las plantas.
  - Hoy seguimos buscando patrones constantemente. La gente de marketing busca patrones en el modo de comprar de las personas, los expertos en seguridad buscan patrones en los ataques que reciben.

Hacer ciencia es buscar patrones

- El trabajo del científico es buscar el sentido en los datos que dispone, para descubrir los patrones que regulan como funcionan las cosas en el mundo real, y crear con ellos teorías que predigan como se comportará un sistema ante nuevas situaciones.

Pero DM no es lo mismo...

- Las técnicas de DM se aplican casi siempre sobre datos que ya fueron almacenados con otro propósito (por ejemplo, registros de ventas de un supermercado)
- DM no tiene influencia en la estrategia de recolección y medición de los datos.
  - Esa es la mayor diferencia con un experimento (con estadística) clásico, donde los datos se miden con una estrategia eficiente que apunta a contestar una pregunta específica.
- Por esto a DM se lo considera un análisis “secundario” de datos

El tamaño sí importa!

- DM trabaja en general sobre datasets grandes
- Se llama *n* al número de registros en un dataset y *p* al número de variables.
- Puede ser grande en *n* o en *p*
- *n* grande: problemas de manejo de los datos
- *p* grande: problemas de validez de los patrones
- Cuando *n* y *p* son chicos se considera al problema como una exploración estadística.

DM y bases de datos

- La comunidad de bases de datos tiende a ver el proceso de DM simplemente como tipos elaborados de queries
  - Como ejemplo, con un query estandar uno puede responder que porcentaje de los pacientes que se hacen un by-pass estuvieron más de 10 días internados en recuperación.
  - Un query no puede responder cuales son las condiciones pre-operatorias relevantes para predecir una recuperación prolongada. Se necesita un modelo de los datos para eso.
- Esta visión llevó a cierta confusión entre DM y queries avanzados.

Algunas aplicaciones comunes

- “Fidelización” de clientes
  - La gente tiende a cambiar de proveedor de telefonía/internet/cable regularmente.
  - Las compañías tratan de detectar cuales son los clientes con mayores posibilidades de dejar el servicio para hacerles una oferta mejor.
  - Es importante no hacerle la oferta a quien no pensaba dejar el servicio!

Algunas aplicaciones comunes

- “Robo” de clientes
  - La gente tiende a no cambiar de proveedor de telefonía/internet/cable por ciertas razones.
  - Las compañías tratan de detectar cuales son esas razones, para modificar sus ofertas hacia los clientes de otras compañías.

Algunas aplicaciones comunes

- Sistemas de recomendación
  - Amazon: Gente que compró … también compró …
  - Gmail: Publicidad “dirigida” basada en el mail que estás leyendo.
  - Netflix prize: Concurso público para desarrollar un sistema de recomendación de películas 10% mejor que el que tenían en 2008/09. Premio U$S 1.000.000. Base pública con 500.000 rankings de películas producidas por usuarios.

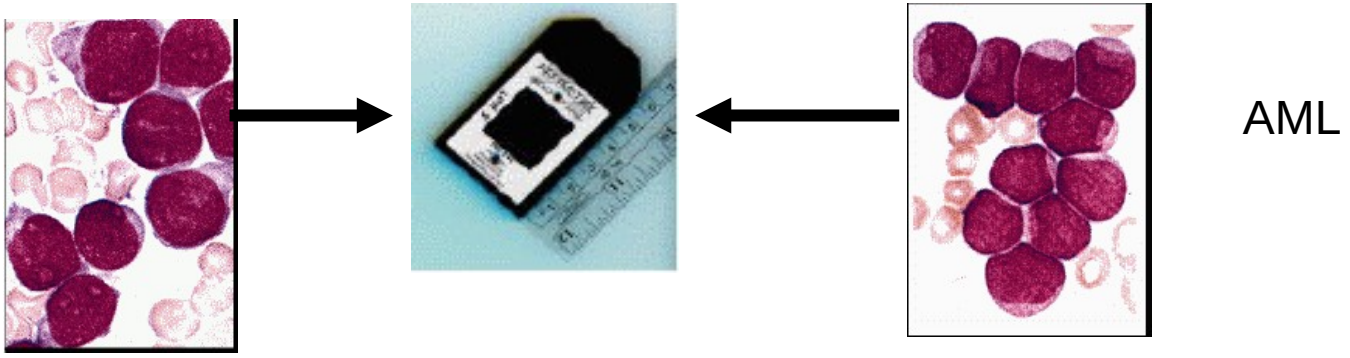
Algunas aplicaciones comunes

- Detección de anomalías
  - Compras fraudulentas con tarjetas/internet
  - Detección de intrusos en cámaras de seguridad

Aplicaciones en ciencia

No son directamente DM, son las mismas técnicas aplicadas como análisis primario.

- Ejemplo: DNA microchips
  - 3 tipos de leucemia. 70 casos. 7000 genes
  - Se puede conocer el tipo de leucemia con este análisis?
  - Cuales son los genes que separan las enfermedades?

ALL  AML

Tareas en DM

- Análisis exploratorio de datos
  - Explorar los datos sin un objetivo particular
  - Herramientas visuales sobre todo
  - Buscar outliers, relaciones simples, etc.
  - Usar proyecciones para datos multidimensionales

Tareas en DM

- Análisis descriptivo
  - Objetivo: describir los datos en forma simple/entendible
  - Estimaciones de densidad: modelar p(x)
  - Clustering: Buscar agrupaciones naturales
  - Relaciones entre variables

  Ejemplo: segmentación de consumidores

Tareas en DM

- Análisis predictivo
  - Se concentra en modelar la relación de todos los datos con una de las variables
  - Clasificación/Regresión

    Ejemplos: Discriminar leucemias en base a DNA

Tareas en DM

- Descubrimiento de patrones/reglas
  - Búsqueda de relaciones en los datos de particular interés. Por ejemplo: la gente que compra asado compra vino.
  - Búsqueda de casos raros (outliers). Ejemplo: detección de fraudes en compras electrónicas.

Tareas en DM

- Recuperación(?) de información
  - Búsqueda de relaciones en los datos que permitan realizar búsquedas de contenidos similares: Sistemas de recomendación.

    Ejemplo: google, amazon

Tópicos de DM

- No vamos a ver DM “comercial”
- Vamos a cubrir sólo algunas etapas del proceso
- En particular las etapas que me parecen más interesantes y que agregan a lo que vimos en ML

Resumen del curso

- Introducción
  - Intro a DM. Repaso lenguaje R. Repaso de ML. TP1
- Pre-procesamiento de datos
  - Visualización. Outliers. Datos faltantes. Proyecciones. Selección de variables. TP2
- Clustering
  - Partición. Jerárquicos. Estabilidad de soluciones. TP3
- Métodos avanzados de clasificación/regresión
  - Ensembles. Bagging y boosting. Random Projections RF.
  - Métodos de kernel. TP4
  - Redes profundas (casi seguro)

Bibliografía

- Principles of Data Mining. Hand et al., 2001
- Data Mining. Witten y Frank, 2005 (WEKA)
- Pattern Recognition and Machine Learning, Bishop, 2006
- Pattern Classification, Duda et al., 2002
- Algunos papers
- R: cran.r-project.org

Hay pdfs de todos los libros y papers.

Método de aprobación

- Asistencia regular a las reuniones virtuales
- 4 trabajos prácticos (en fecha). Puntaje por opcionales y por fecha de entrega.
- LCC: buscar un dataset y analizarlo
- Doctorado: un trabajo profundo sobre algún tema relacionado al curso
- Ambos: dar una charla de 20' sobre un paper

(temas y papers a convenir)