Business Intelligence
Data Mining
Big Data
Machine Learning

Extracto del curso brindado en el CPCECABA

Fraquelli

¿Que es Datamining ?

Conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Fraquelli

¿Porqué hacer minería de datos?

- Se recolectan y almacenan gran cantidad de datos
  - Internet, celulares Comercios Bancos, tarjetas de crédito, CRM, Redes Sociales…
- Las computadoras son más económicas y potentes
- La presión competitiva de los negocios
  - Proveer mejores servicios y más focalizados en las necesidades de cada cliente

Fraquelli

Ejes Principales de la minería de datos

Fraquelli

¿Cual estadística?

Descriptiva = obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso: MINIMOS / MAXIMOS / PROMEDIOS / CUARTILES / MEDIAS / DISTRIBUCIÓN

Inferencial = Probabilidades / comprende los métodos y procedimientos que por medio de la inducción determina propiedades de una población. Obtener conclusiones útiles para hacer deducciones

Fraquelli

¿Como Funciona?

Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias.

El algoritmo usa los resultados de este análisis en un gran número de iteraciones para determinar los parámetros óptimos para crear el modelo de minería de datos

A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

¿Como Funciona?

El modelo de minería de datos que crea un algoritmo a partir de los datos puede tomar diversas formas, incluyendo:

•Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.

•Un árbol de decisión que predice un resultado y que describe cómo afectan a este los distintos criterios.

•Un modelo matemático que predice las ventas.

•Un conjunto de reglas que describen cómo se agrupan los productos en una transacción, y las probabilidades de que dichos productos se adquieran juntos. *** Cervezas y Pañales *** 1,67 +/- 20cm

¿Como Elegir el algoritmo correcto?

La elección del mejor algoritmo para una tarea analítica específica puede ser un desafío.

Aunque puede usar diferentes algoritmos para realizar la misma tarea, cada uno de ellos genera un resultado diferente, y algunos pueden generar más de un tipo de resultado

Algoritmos mas comunes

- **Algoritmos de clasificación**, que predicen una o más variables discretas, basándose en los demás atributos del conjunto de datos.

- **Algoritmos de regresión**, que predicen una o más variables numéricas continuas, como pérdidas o ganancias, basándose en otros atributos del conjunto de datos.

- **Algoritmos de segmentación**, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.

Algoritmos más comunes

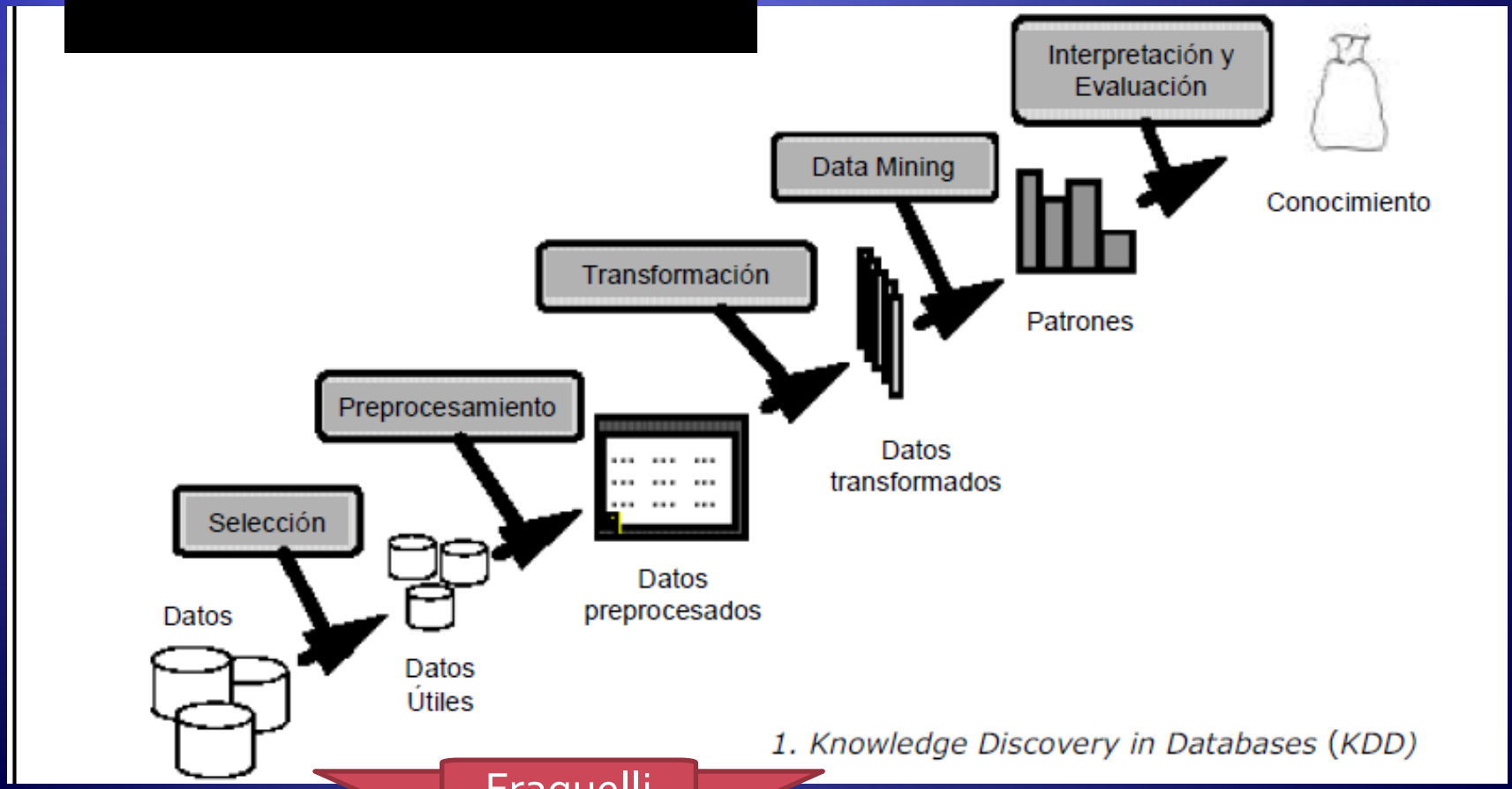
Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra.

Algoritmos de análisis de secuencias, resumen las secuencias frecuentes o episodios en los datos, como una serie de clics en un sitio web o una serie de eventos de registro que preceden al mantenimiento del equipo.

Algoritmos combinados

Los analistas experimentados usarán a veces un algoritmo para determinar las entradas más eficaces (es decir, variables) y luego aplicarán un algoritmo diferente para predecir un resultado concreto basado en esos datos

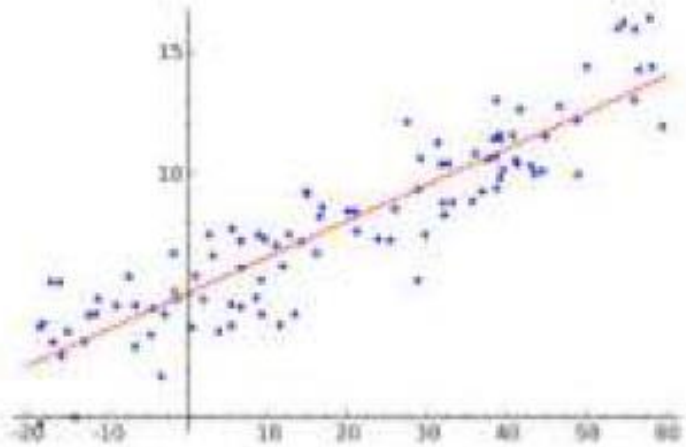
¿Cuáles son las etapas del proceso de descubrimiento de conocimiento a partir de los datos?

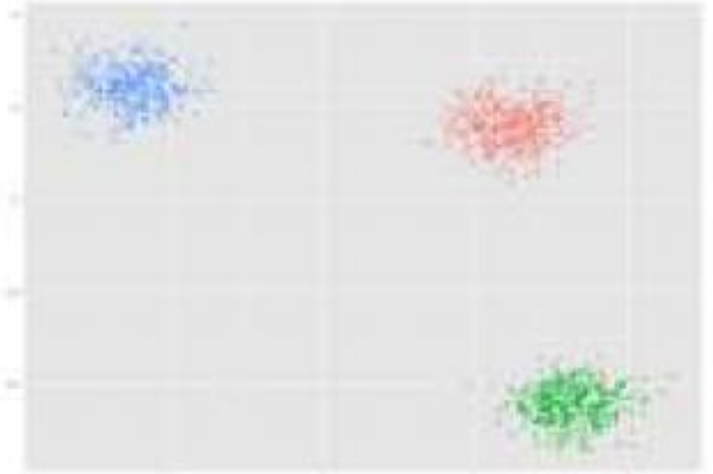


Tareas Data Mining

- **Método predictivo**
  - Utilizan algunas variables para predecir valores futuros o desconocidos de otras variables

- **Método descriptivo**
  - Encuentran patrones que describen los datos, que son interpretables por los humanos.





Fraquelli

- **Método predictivo**
  - Clasificación
  - Regresión
  - Detección de desvíos

- **Método descriptivo**
  - Clustering
  - Reglas de asociación
  - Patrones secuenciales

1.¿ Qué es Data Mining?

- El Data Mining es un proceso que nos permite descubrir o incrementar el conocimiento que se posee de una cierta área a partir de la aplicación de una serie de técnicas, haciendo uso de las tecnologías desarrolladas a partir de las bases de datos, estadística y aprendizaje automático.
- Para algunos esta definición debe completarse con la adición de que tal cometido se logra a partir del análisis de grandes volúmenes de datos almacenados en sistemas informáticos (Big Data).

Fraquelli

2.Describir los pasos del proceso de descubrimiento de conocimiento KDD.

- Las etapas del proceso de descubrimiento de conocimiento o Data Mining a partir de los datos de una o varias base de datos son:
  - Selección,
  - Pre-procesamiento,
  - Transformación,
  - Data Mining propiamente dicho, e
  - Interpretación y Evaluación.

Fraquelli

3.Discutir si cada una de las siguientes actividades es o no una tarea de data mining:

a. Dividir los clientes de una compañía de acuerdo a su género.

b. Dividir los clientes de una compañía de acuerdo a su rentabilidad.

c. Calcular el total de ventas de una compañía

d. Clasificar una base de datos de estudiantes basado en los números de identificación.

e. Predecir los resultados de arrojar un par de dados.

f. Predecir el precio futuro de las acciones de una compañía usando registros históricos.

g. Monitorear los latidos del corazón de un paciente para buscar anomalías.

h. Monitorear ondas sísmicas para detectar actividades de terremotos.

i. Extraer las frecuencias de una onda de sonido.

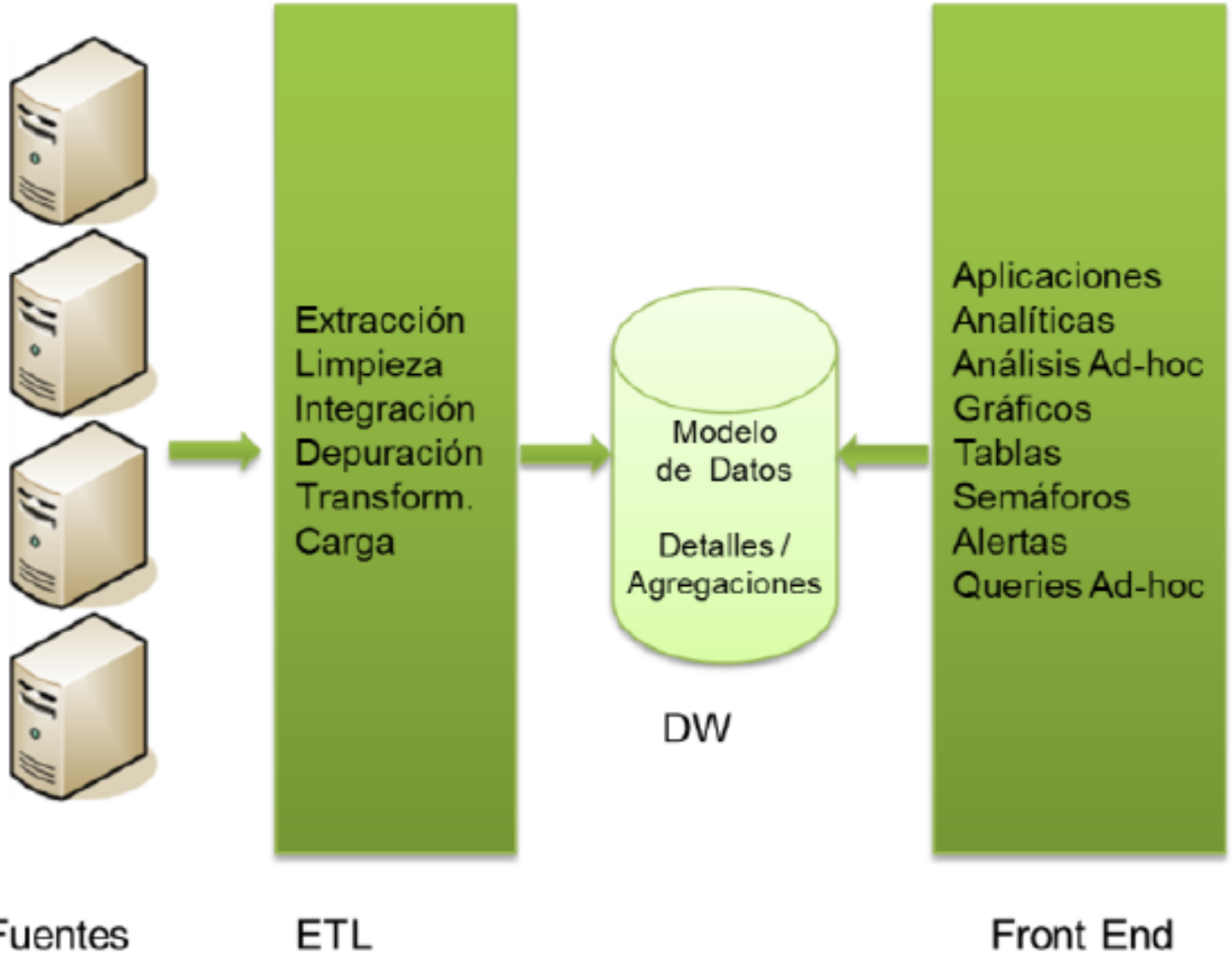
4.Suponer que usted tiene un empleo como consultor en data mining para una compañía de un buscador de Internet. Describir cómo puede data mining ayudar a la compañía por medio de ejemplos específicos de cómo se pueden aplicar técnicas como Clustering, Clasificación, Descubrimiento de Reglas de Asociación y Detección de Anomalías.

- Clustering: determinar tipos de grupos de usuarios que acceden a sitios de compra de productos por Internet a partir de algunas características particulares básicas (edad, género, país, etc.): jóvenes de países del tercer mundo, personas adultas solteras, etc.
  - De esta manera se pueden mejorar las estrategias de marketing para diferentes productos a ser ofertados a través de Internet.
- Clasificación: determinar la clase de producto de interés de cada usuario del buscador en base a sus datos básicos.
  - De esta manera se pueden recomendar opciones de compra a los diferentes usuarios del buscador. Netflix? Spotify? Youtube? Tinder? Campañas Políticas...?

Continua…

- Reglas de asociación: en base a los registros históricos de las últimas visitas a páginas web determinar las posibles recomendaciones a realizar en base a las búsquedas realizadas en la actualidad.
  - Otras personas que vieron XXX también consultaron ZZZ
- Detección de anomalías: detectar hackers en base a la secuencia de páginas visitadas.
  - Han accedido a tu dispositivo desde una IP no registrada previamente…

DW – Un Esquema Genérico y Simple



Fraquelli


Semi-Structured
Unstructured data
Petabytes
analysis
content
BIG
DATA
Volume
Useful
Metadata
AMP
Structured
People driven
Decision making
Data
Zettabyte
Unstructured
Framework
Smart content
database
IT
VA
Concept extraction
Text analytics
Semantic Metadata

Fraquelli

Las tres V del Big Data

Volumen

- Gigantesca cantidad de datos
  - PETABYTES

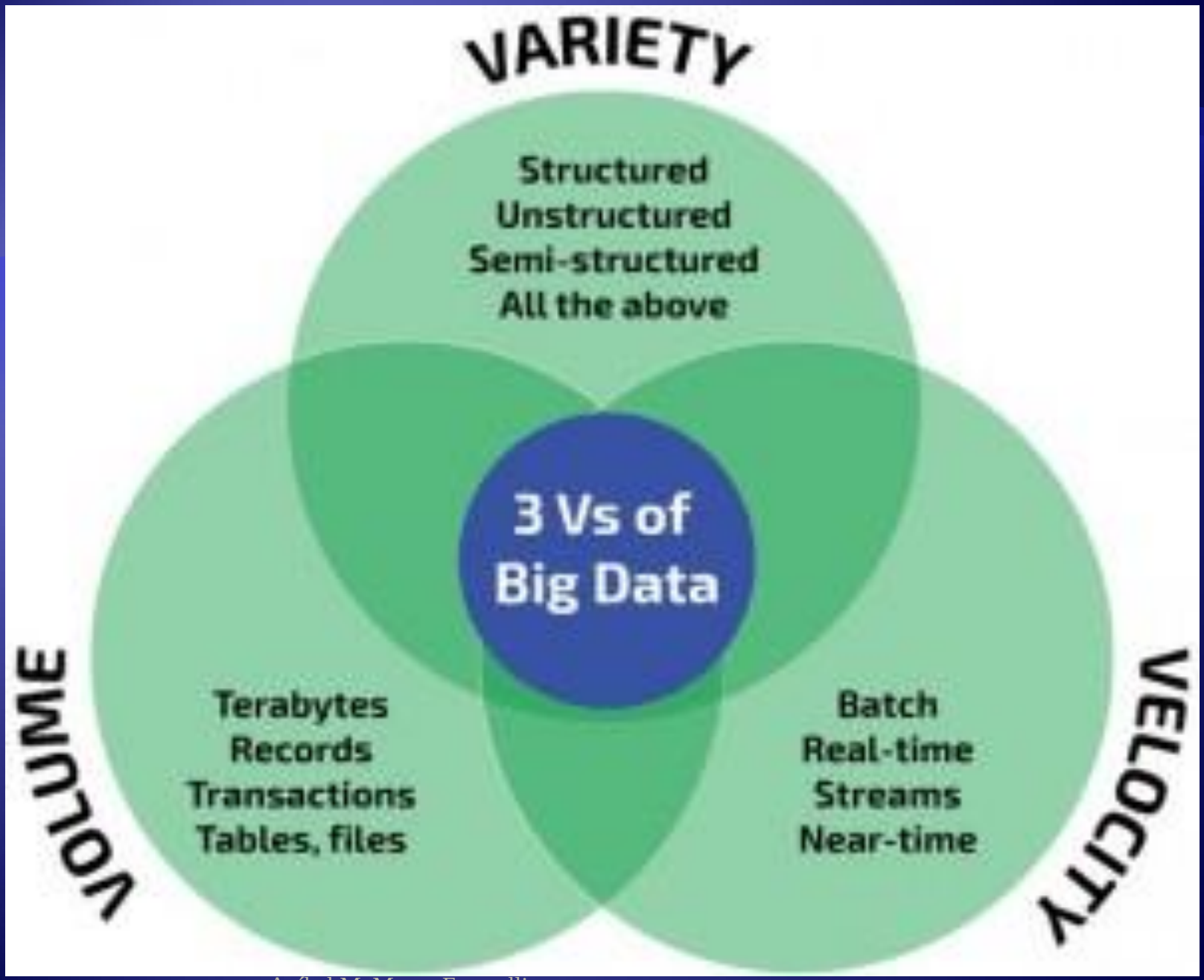
Velocidad

- Problemas a considerar:
  - Vel en que la info está disponible para analisis
  - Velocidad en que podemos actuar luego de procesar

Variedad

- Diferentes tipos de Data:
  - Estructurada (20 %)
  - Semi – Estructurada(10%)
  - No Estructurada(70%)

Fraquelli


VARIETY
Structured
Unstructured
Semi-structured
All the above
3 Vs of
Big Data
VOLUME
Terabytes
Records
Transactions
Tables, files
VELOCITY
Batch
Real-time
Streams
Near-time

Unidades de almacenamiento (tradicionales)

Nombre	Símbolo	Binario	Número de bytes	Equivale
kilobyte	KB	2^10	1.024	=
megabyte	MB	2^20	1.048.576	1.024KB
gigabyte	GB	2^30	1.073.741.824	1.024MB
terabyte	TB	2^40	1.099.511.627.776	1.024GB
petabyte	PB	2^50	1.125.899.906.842.624	1.024TB
exabyte	EB	2^60	1.152.921.504.606.846.976	1.024PB
zettabyte	ZB	2^70	1.180.591.620.717.411.303.424	1.024EB
yottabyte	YB	2^80	1.208.925.819.614.629.174.706.176	1.024ZB

Big Data Analitycs

- Examinar grandes cantidades de data en tiempos razonables
- Información adecuada y oportuna
- Identificación de patrones escondidos o correlaciones desconocidas en datos estructurados y no estructurados
- Ventajas competitivas por patrones o mapas de calor en segmentos de Mercado
  - Mejores decisiones de Negocio: Estrategicas y operacionales
- Eficiencia en Marketing, satisfacción del cliente -> Aumento del Revenue (por poco stock, mejor logistica, menor cantidad empleados para procesos operativos).

Fraquelli

Machine Learning

- Convertir "Datos" en "Conocimiento"
- Test de Turing (Lenguaje Natural)
- Inteligencia Artificial
- Inducción de Conocimiento
- Sistemas: Ayudante – Colega – Autómata
- Aprendizaje:
  - Supervisado / No supervisado / semi supervisado
  - Por refuerzo / transducción / Multi tarea (reaprendizaje)

Fraquelli