



ORDÓÑEZ ORDÓÑEZ & ASOCIADOS LTDA.  
Asesores en Tecnología Informática

# Minería de Datos

## Cómo abordar un proyecto en su Organización

ACJS

María Esther Ordóñez O.  
Febrero 2010

[mordonez@ordonezasesores.com.co](mailto:mordonez@ordonezasesores.com.co)



# Agenda

- **Tecnologías Estratégicas - Gartner 2010**
- La nueva generación de BI
- Minería de Datos y BI
- Metodología CRISP-DM
- Conclusiones y Recomendaciones

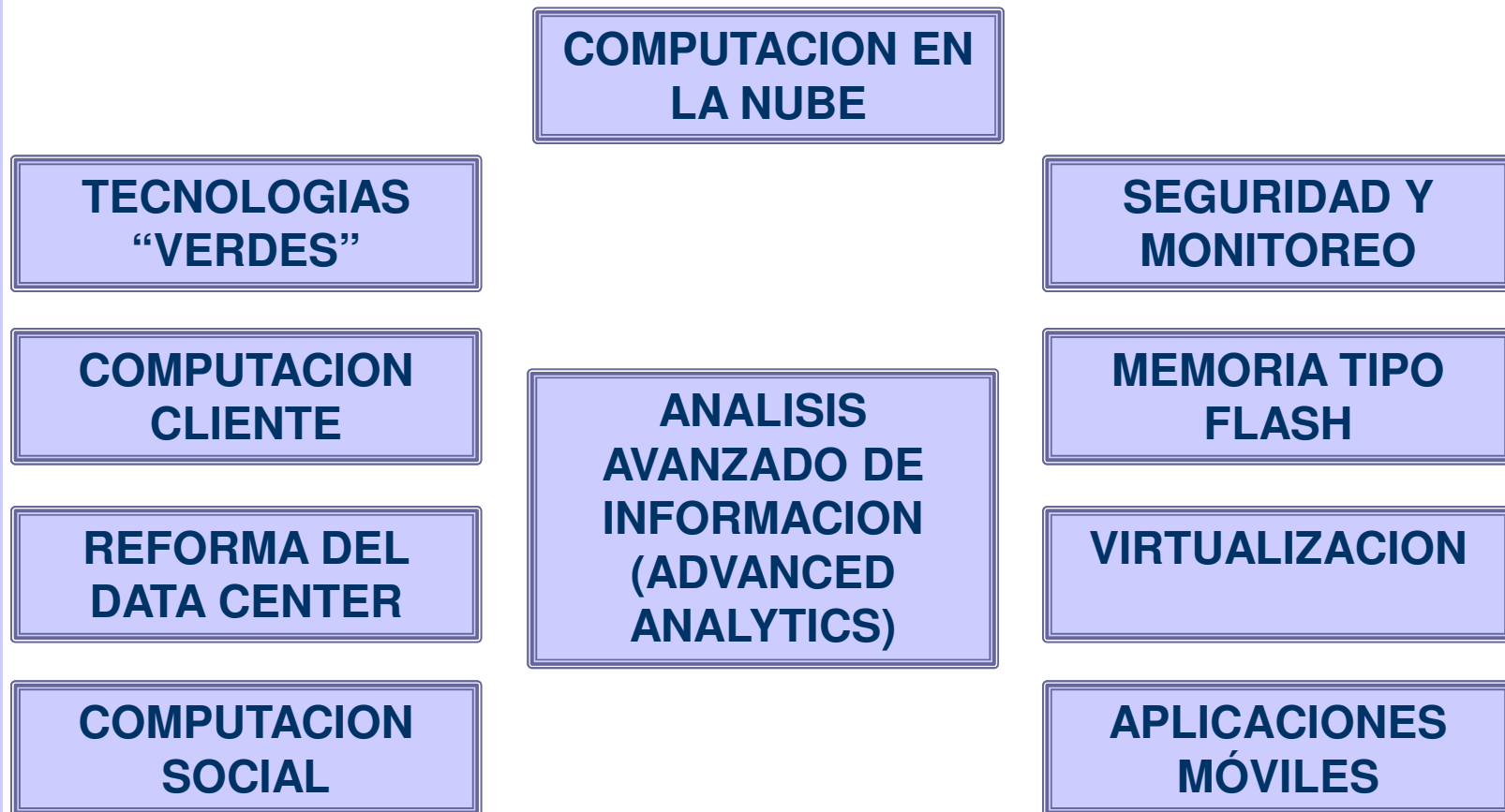


# Tecnologías estratégicas 2010

- Grupo Gartner - Octubre de 2009
- Impactarán significativamente los negocios en los siguientes 3 años
  - Mayor potencial de “trastornar ” IT o Negocio
  - Mayor requerimiento de inversión
  - Implica riesgo no adoptarlas a tiempo
- Impactan los planes, programas e iniciativas a largo plazo, y son estratégicas
  - Son maduras y de amplia utilización
  - Generan ventajas estratégicas por adopción temprana



# Tecnologías Estratégicas 2010





# Advanced Analytics

“.....The new step is to provide **simulation,**  
**prediction, optimization** and other  
analytics, **not simply information,** to  
**empower even more decision flexibility at**  
**the time and place** of every business  
process action. The new step looks into the  
future, predicting what can or will happen....”

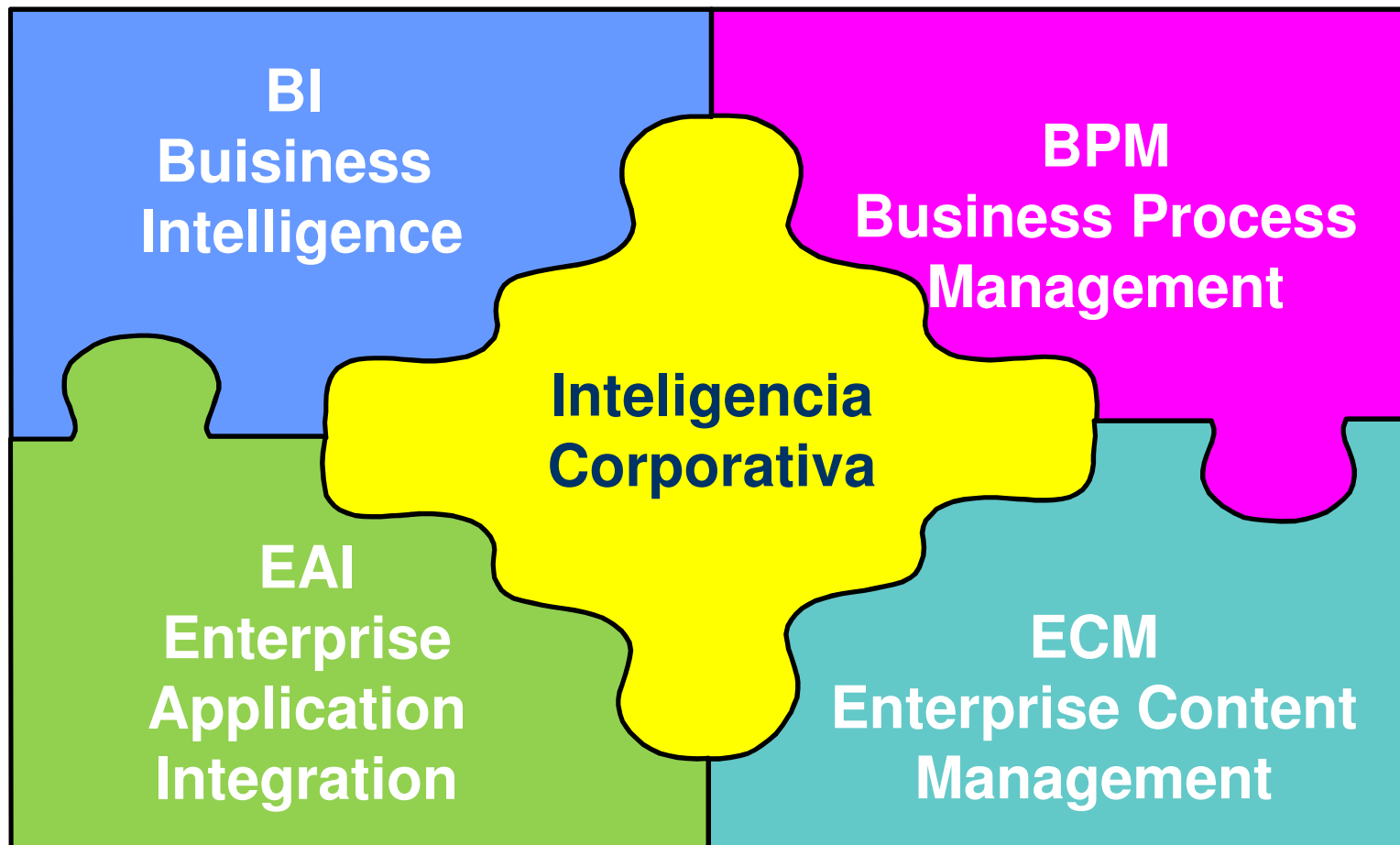


# Agenda

- Tecnologías Estratégicas - Gartner 2010
- **La nueva generación de BI**
- Minería de Datos y BI
- Metodología CRISP-DM
- Conclusiones y Recomendaciones

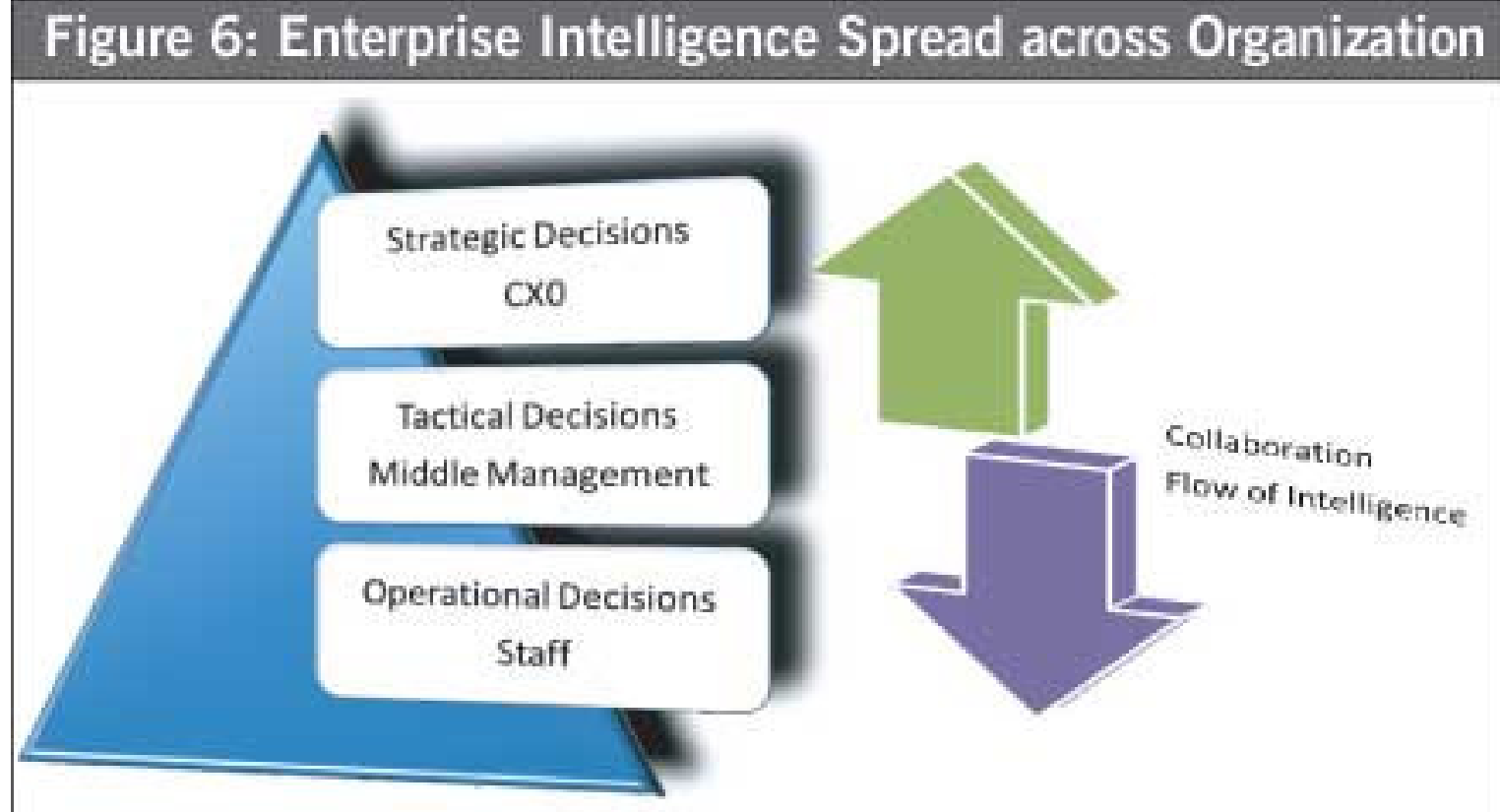


# La Estrategia...





# El estado ideal.....



Fuente: DMReview.com, Mayo 13 2008



# Agenda

- Tecnologías Estratégicas - Gartner 2010
- La nueva generación de BI
- **Minería de Datos y BI**
- Metodología CRISP-DM
- Conclusiones y Recomendaciones

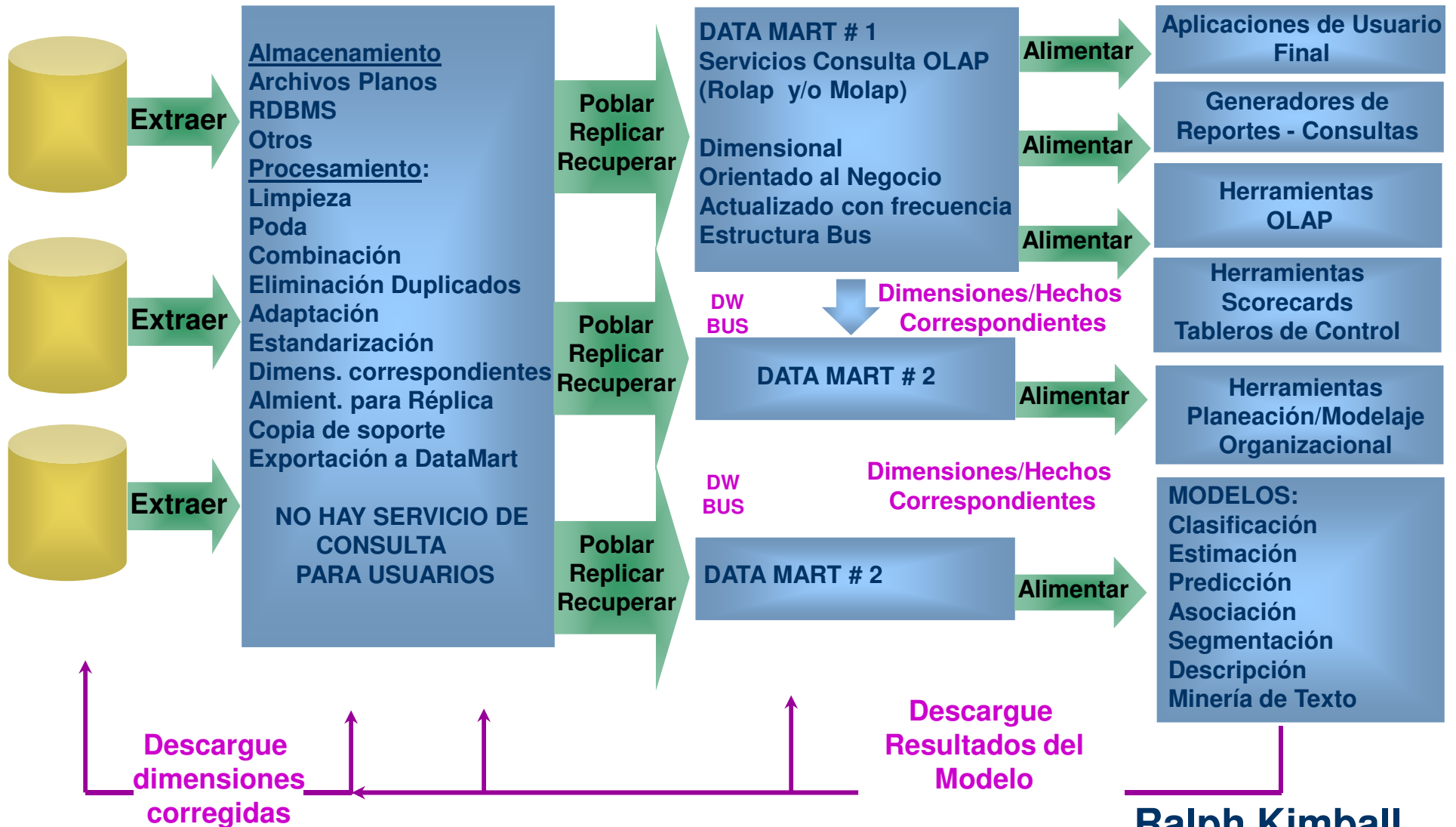


## Sistemas Fuente (Source Systems)

## Area de Preparación de Datos - ETL (Data Staging Area)

## Servidor de Presentación BODEGA DE DATOS

## Usuario Final Acceso a Datos



Ralph Kimball



# Toma de Decisiones.....

## Tres Capas

Gráfica,  
Datos Abstraídos

**Monitoreo**  
Tableros de Control – KPI - Alertas



Sumarizada,  
Datos Dimensionales

**Análisis**  
Dimensiones – Jerarquías – Slice/Dice



Detallada,  
Datos Operacionales

**Reporte**  
Reportes administrativos y Operativos



**Planeación**  
Planes – Modelos - Estimaciones



	RECOLECCION DE DATOS	ACCESO A DATOS	BODEGAS DE DATOS Y DSS	DATA MINING
PREGUNTA DE NEGOCIO	Cuáles fueron los ingresos el año anterior	Cuáles fueron las ventas de Cali en Marzo	Análisis de ventas a diferentes niveles temporal y geográfico	Cuáles serán las ventas en períodos futuros
TECNOLOGIA HABILITADORA	Computadores, cintas, discos	Bases de Datos relacionales y SQL	Bodegas de Datos, Bases de Datos multidimensionales, OLAP	Algoritmos avanzados, multiprocesador, bases de datos masivas
CARACTERISTICAS	Resúmenes de datos pasados y estáticos	Datos dinámicos del pasado a nivel detallado	Datos dinámicos del pasado a múltiples niveles	Información proactiva - predictiva

## Desarrollo Histórico de la Minería de Datos



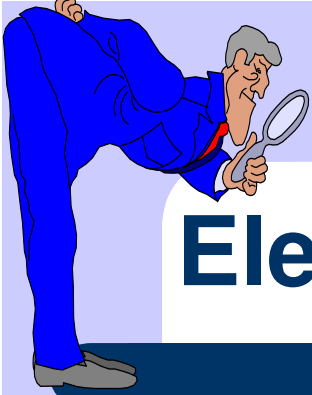
# Minería de Datos

Exploración y Análisis de datos, utilizando métodos automáticos o semi-automáticos, con el objeto de descubrir patrones no evidentes, significativos y reglas de comportamiento



# Minería de Datos

Es un *Proceso de Negocio* cuyo objetivo es encontrar patrones en los datos, no evidentes y significativos, que generen conocimiento e ideas de cómo conducir el negocio de una manera más eficiente y eficaz



# Elementos Eticos y Legales

- Pueden verse como métodos de discriminación
- Es necesario tener en cuenta las condiciones bajo las cuales se recoge la información - Las personas deben ser informadas de los objetivos del proceso
- Debe existir una explicación de negocio asociada a una decisión tomada con base en DM



# Agenda

- Tecnologías Estratégicas - Gartner 2010
- La nueva generación de BI
- Minería de Datos y BI
- **Metodología CRISP-DM**
- Conclusiones y Recomendaciones



# Metodología CRISP-DM

CRISP-DM 1.0

Cross-Industry Standard Process for  
Data Mining

**Step-by-step data mining guide  
2000**

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber  
(NCR), Thomas Khabaza (SPSS), Thomas Reinartz  
(DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth  
(DaimlerChrysler)



**Entendimiento  
del negocio**

**Determinar  
Objetivos de  
Negocio**



**Verificar  
Situación  
Actual**



**Determinar  
Objetivos  
de Minería**



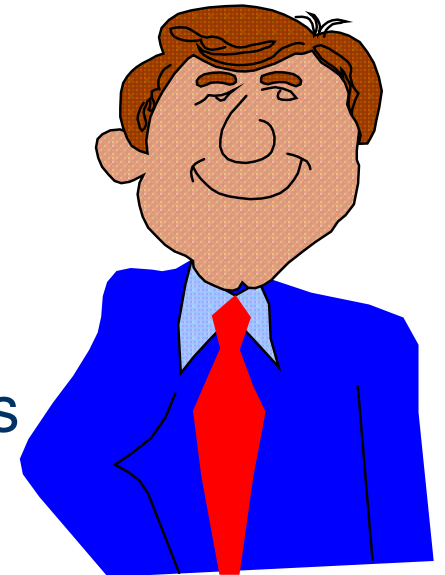
**Producir  
Plan del  
Proyecto**





# Objetivos Generales DM

- **Conocimiento y Fidelización de clientes**
  - Segmentación
  - Cross-sell y Up-sell
  - Manejo del ciclo de vida del cliente
  - Conocimiento Comunitario
  - Identificar perfiles deseables para nuevos negocios
  - Optimización de campañas de Mercadeo
  - Manejo de Deserción de clientes





# Objetivos Generales DM

- Detección de Fraudes
  - Identificación de patrones de comportamiento normal / fraudulento





# Objetivos Generales DM

- Eficiencia de Procesos
  - Se aplican técnicas de DM a casos ya resueltos para determinar reglas
  - Control estadístico de procesos de manufactura



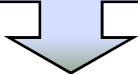


# Objetivos Generales DM

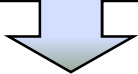
- Mejor definición de productos, combos
- Administración de espacios en los almacenes de retail
- Aislamiento de Fallas en redes de servicios
- Optimización en la planeación y localización de recursos
- Nuevos Negocios (brokers de información)
- Mejorar Servicios (Text Mining)



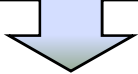
Determinar  
Objetivos de  
Negocio



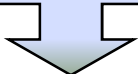
Verificar  
Situación  
Actual



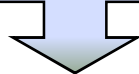
Determinar  
Objetivos de  
Minería



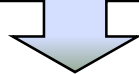
Producir  
Plan del  
Proyecto



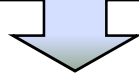
Recolectar  
Datos  
Iniciales



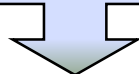
Describir  
Datos



Explorar  
Datos



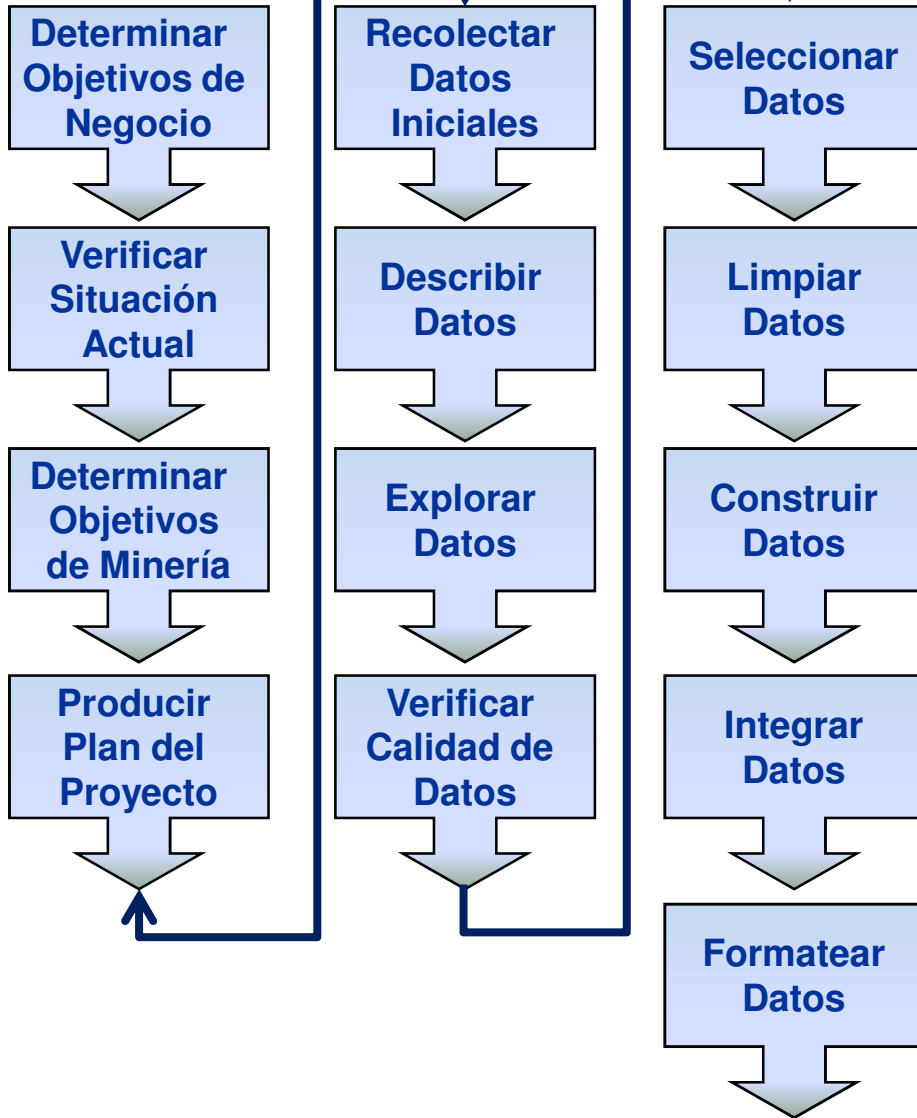
Verificar  
Calidad de  
Datos





# Entendimiento de los Datos

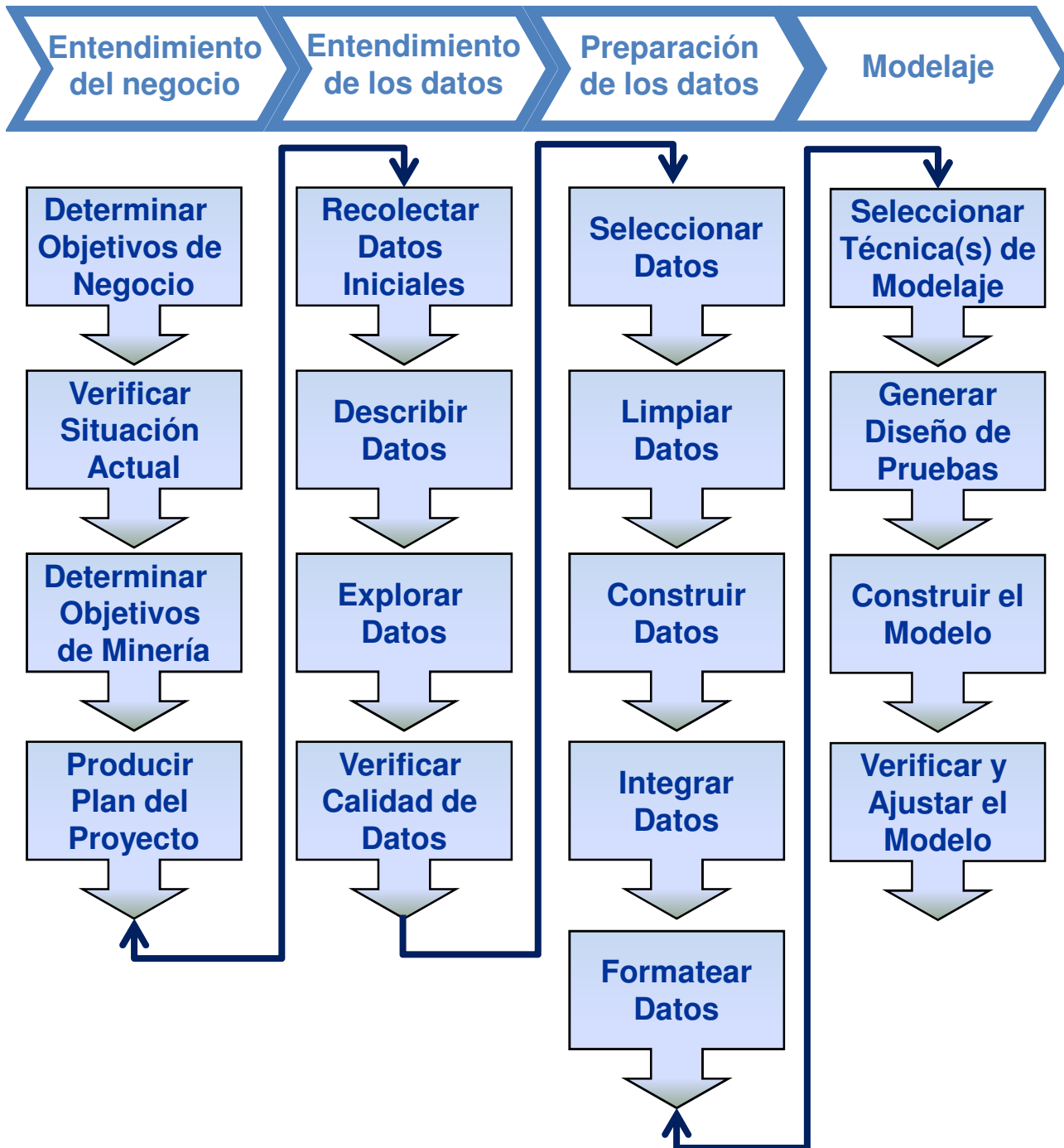
- Herramientas de ETLC
- Medidas Estadísticas (media, sd etc.)
- Análisis Multivariado






# Preparación de los Datos

- Herramientas de ETL
- Escoger atributos
- Incluir medidas derivadas o calculadas
- Incluir casos “positivos” y “negativos”
- Dividir en conjunto de Entrenamiento – Prueba - Evaluación
- Oversamplig





# Minería de Datos - Tareas

- Clasificación
  - Estimación
  - Predicción / Forecasting
  - Asociación / Grupo por afinidad
  - Segmentación / Clustering
  - Descripción y Perfilación
  - Análisis Textual
- Dirigido
- No Dirigido
- 



# Clasificación

- Establecer una o más variables discretas de un objeto, con base en otros atributos del conjunto de datos - Analizar características de un nuevo objeto y asignarlo a una clase particular predefinida

- Clasificar solicitud de crédito en riesgo alto - medio – bajo
- Determinar qué teléfonos corresponde a máquinas de fax
- Identificar Reclamos de Seguro fraudulentos
- Clasificar a una persona como potencial “respondedor” a oferta
- Clasifica a un cliente dentro de un perfil particular





# Estimación

- Establecer el valor de una variable continua, los resultados pueden ser ORDENADOS – Similar a Clasificación

- **Estimar el valor del ingreso total de un grupo familiar**
- **Determinar probabilidad de que una transacción sea fraudulenta**
- **Estimar número de hijos en un grupo familiar**
- **Estimar el valor del ciclo de vida de un cliente**
- **Estimar probabilidad con que una persona responde a campaña**





# Predicción

- Similar a clasificación o estimación, sólo que se refiere a identificar un comportamiento o valor estimado futuro

- **Predecir qué clientes desertarán en los siguientes 6 meses**
- **Predecir el monto de saldo transferido si un prospecto de TC acepta la oferta de transferencia**
- **Predecir qué suscriptores de teléfonos ordenarán servicios de valor agregado**





# Asociación/Agrupamiento Afinidad

- Detectar eventos que ocurren de manera simultánea

- Un cliente que compra cerveza, compra pañales con probabilidad P1
- Un cliente que compra Pizza, compra Vino con probabilidad P1
- Un cliente que compra Vino, compra Pizza con probabilidad P2





# Segmentación/Clustering

- Dividir población heterogénea en subgrupos más homogéneos (segmentos o clusters)

- **Establecer los diferentes segmentos en que pueden organizarse los clientes de un negocio particular**





# Descripción

- Describir un comportamiento en una base de datos compleja para aumentar el conocimiento y entendimiento sobre gente, productos, procesos etc.  
– Visualización - Diferenciación

- Establecer que el porcentaje de mujeres que adquieren un determinado producto es mayor que el de hombres
- Establecer que personas de ciertas características apoyan a un candidato liberal o conservador





# Análisis Textual

- Convierte información desestructurada en información estructurada – Análisis de Términos

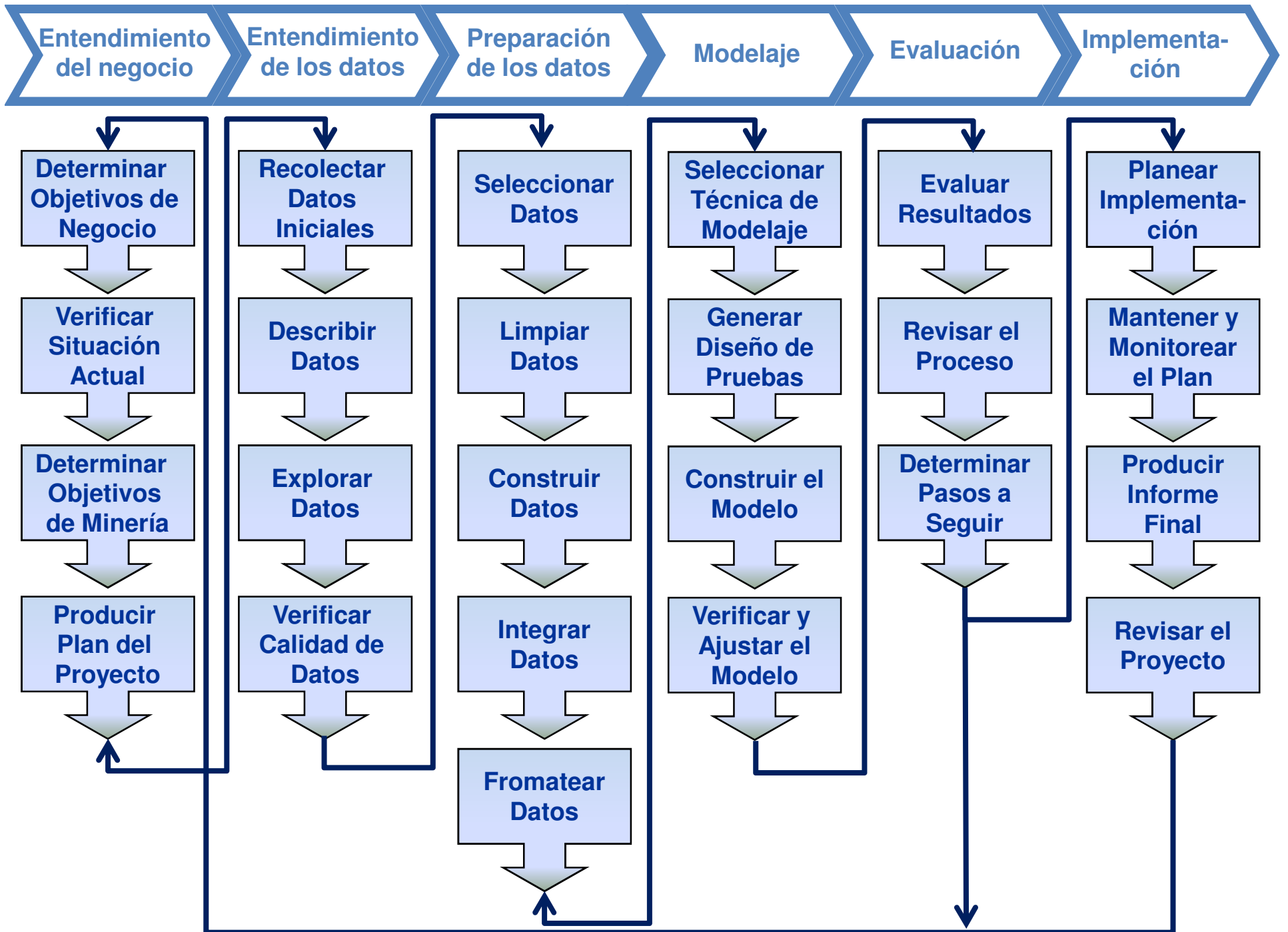
- **Retroalimentación en datos de Call Center**
- **Análisis de correos electrónicos**
- **Clasificación de textos por temas**
- **Análisis de reportes (policiales, de mantenimiento etc)**





# Técnicas

- Árboles de Decisión
- Redes Neuronales
- Clustering
- Reglas de Asociación
- Link Analysis
- Razonamiento Basado en Memoria
- Modelos de regresión lineal y logística
- Text Mining





# Agenda

- Tecnologías Estratégicas - Gartner 2010
- La nueva generación de BI
- Minería de Datos y BI
- Metodología CRISP-DM
- **Conclusiones y Recomendaciones**



# Conclusiones y Recomendaciones

Aunque DM utiliza tecnología avanzada, un proyecto NO DEBE ser desarrollado únicamente por expertos en tecnología.

DM aporta valor únicamente si sus resultados se ponen a servicio del negocio



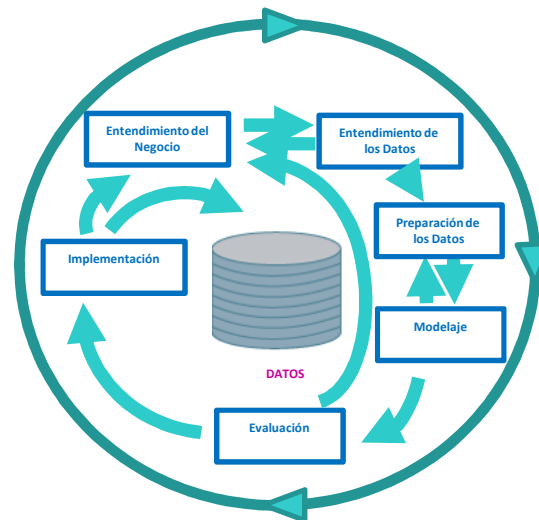
# Conclusiones y Recomendaciones

DM es un proceso INTERACTIVO e  
ITERATIVO



# Conclusiones y Recomendaciones

DM NO consiste únicamente en algoritmos avanzados de análisis de datos





# Conclusiones y Recomendaciones

Aunque NO ES indispensable contar con una DWH para DM, SI ES deseable.

La Bodega debe cumplir el requisito de manejar el **MAXIMO NIVEL DE DETALLE**



# Conclusiones y Recomendaciones

Aunque DM saca provecho de grandes volúmenes de datos, se han obtenido resultados útiles con base en conjuntos de datos de tamaño mediano o pequeño (cientos o miles de registros)



# Conclusiones y Recomendaciones

La disponibilidad, relevancia y calidad de los datos es fundamental para DM



# Conclusiones y Recomendaciones

Una buena estrategia para empezar es desarrollar un prototipo utilizando herramientas libres (WEKA p.e.), pero teniendo en cuenta TODO el ciclo metodológico