MINISTERIO DEL TRABAJO

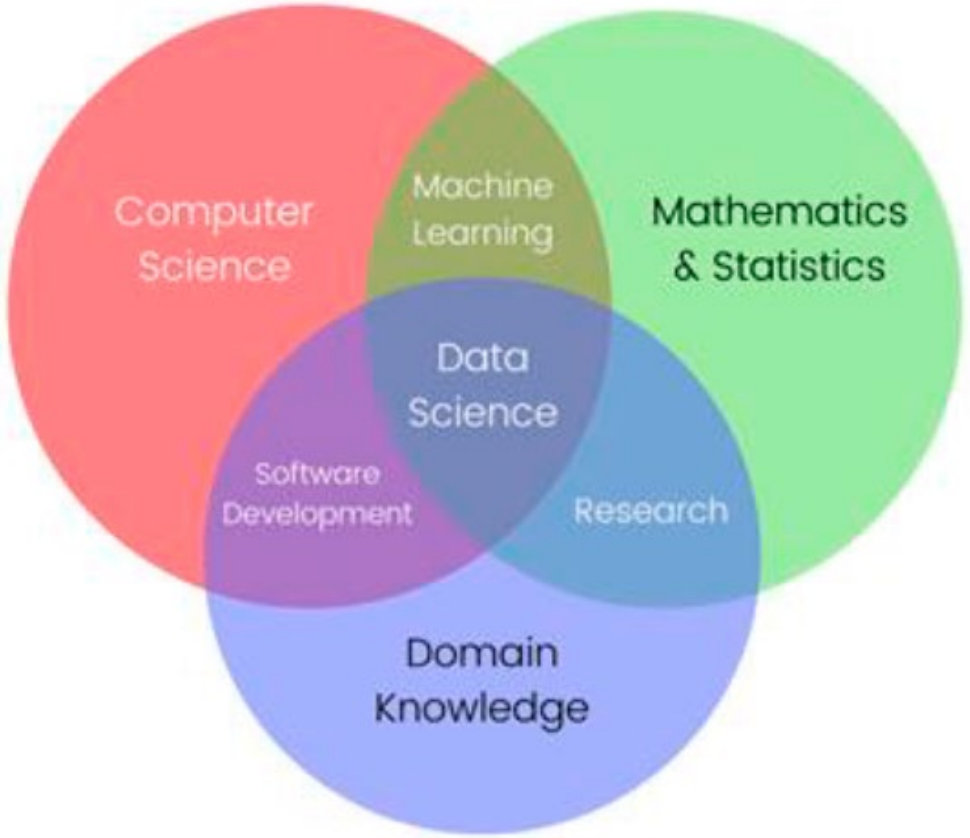
# Ciencia de Datos (Metodologías)

Centro de Servicios y Gestión Empresarial
SENA Regional Antioquia

# Ciencia de Datos

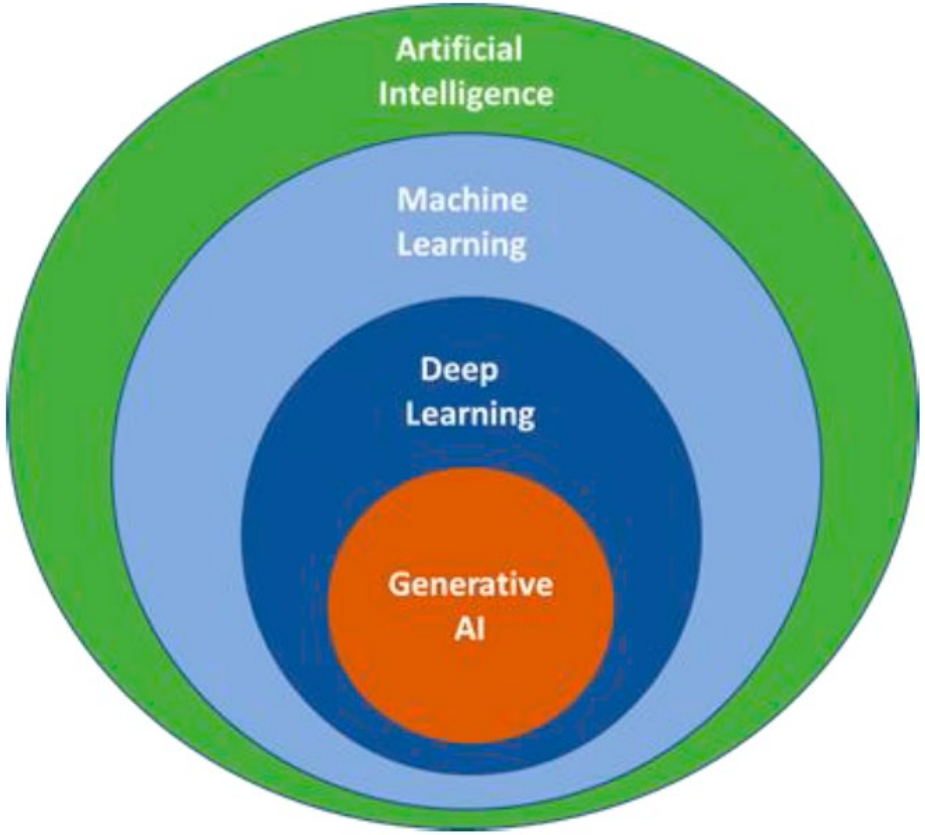
# Ciencia de Datos

La ciencia de datos consiste en extraer conocimientos e ideas de los datos, encontrar patrones invisibles, obtener información significativa y hacer negocios. información significativa y decisiones empresariales



# Inteligencia Artificial

La inteligencia artificial (IA) se refiere a la simulación de procesos de inteligencia humana por parte de máquinas, especialmente sistemas informáticos. Es un campo interdisciplinario que combina la informática, la ciencia de datos, la psicología cognitiva y otros campos para crear sistemas capaces de realizar tareas que normalmente requerirían la intervención humana y el uso del razonamiento humano.



# Metodologías

# Metodología

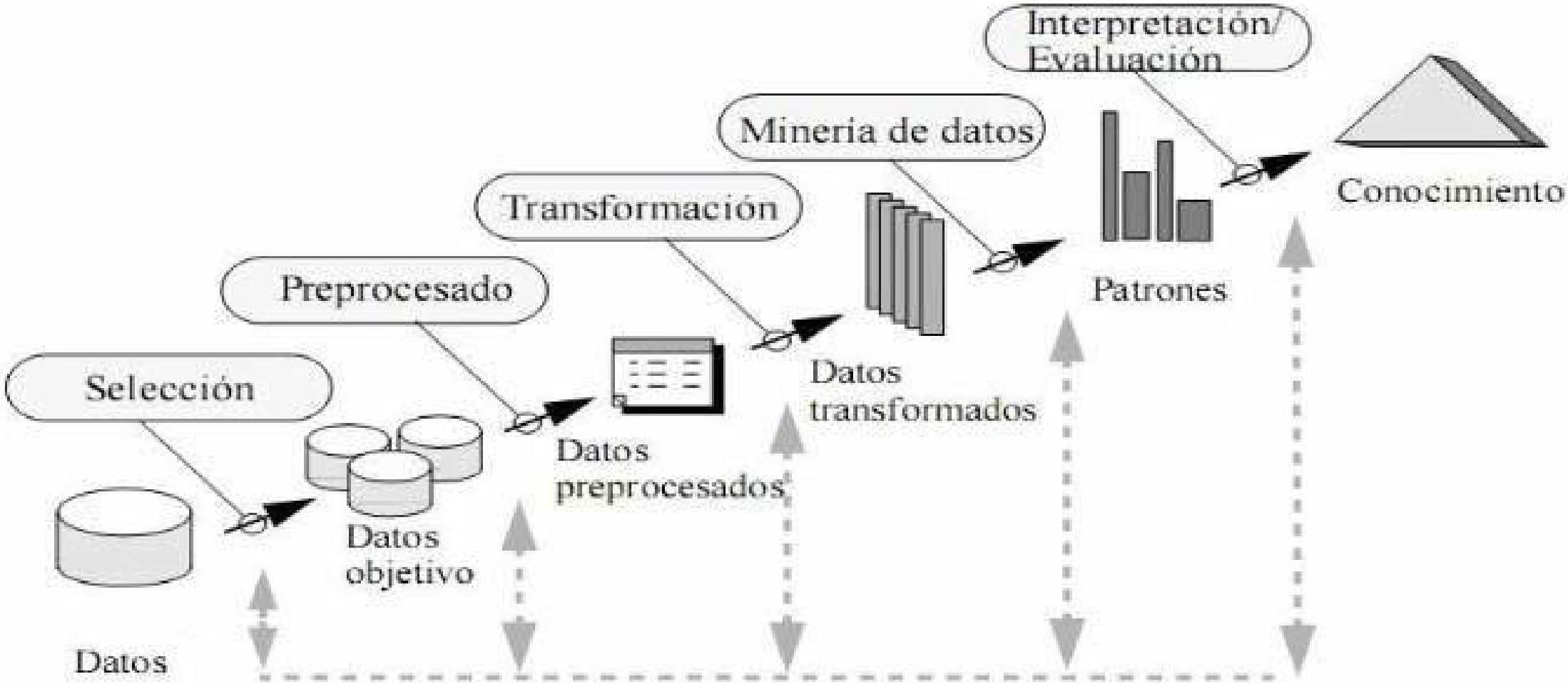
Una **metodología** es un conjunto de métodos, técnicas, prácticas y enfoques sistemáticos que se utilizan para llevar a cabo un proceso o alcanzar un objetivo específico.

En términos simples, es un enfoque organizado y estructurado para abordar tareas, problemas o proyectos de manera eficiente y efectiva.

Las metodologías se aplican en una variedad de campos, como la investigación, el desarrollo de software, la gestión de proyectos, la educación y más.

# Tipos de Metodologías

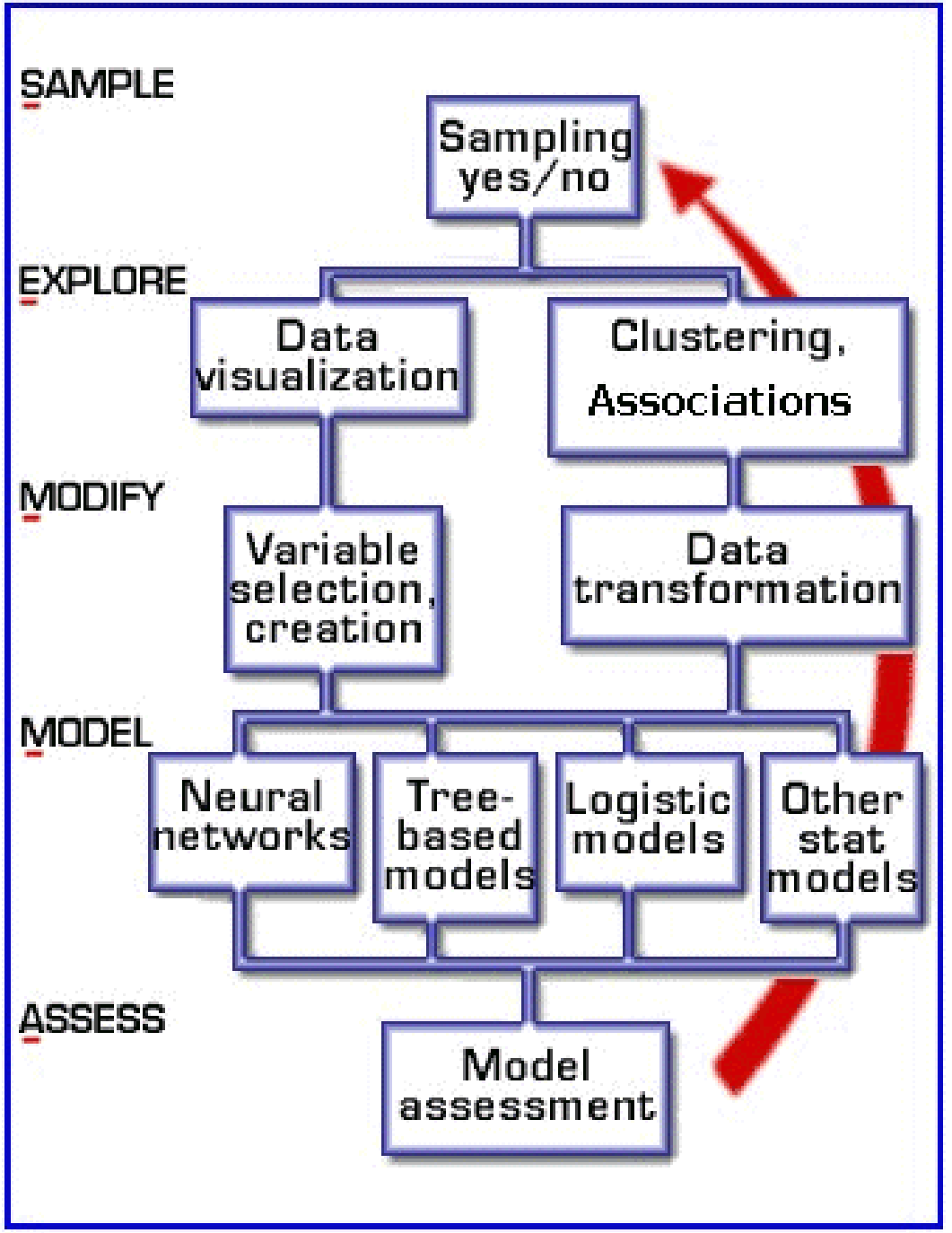
- KDD
- SEMMA
- Catalyst ó P3TQ
- CRISP-DM




**Modelo KDD (*Knowledge Discovery in Databases*)**

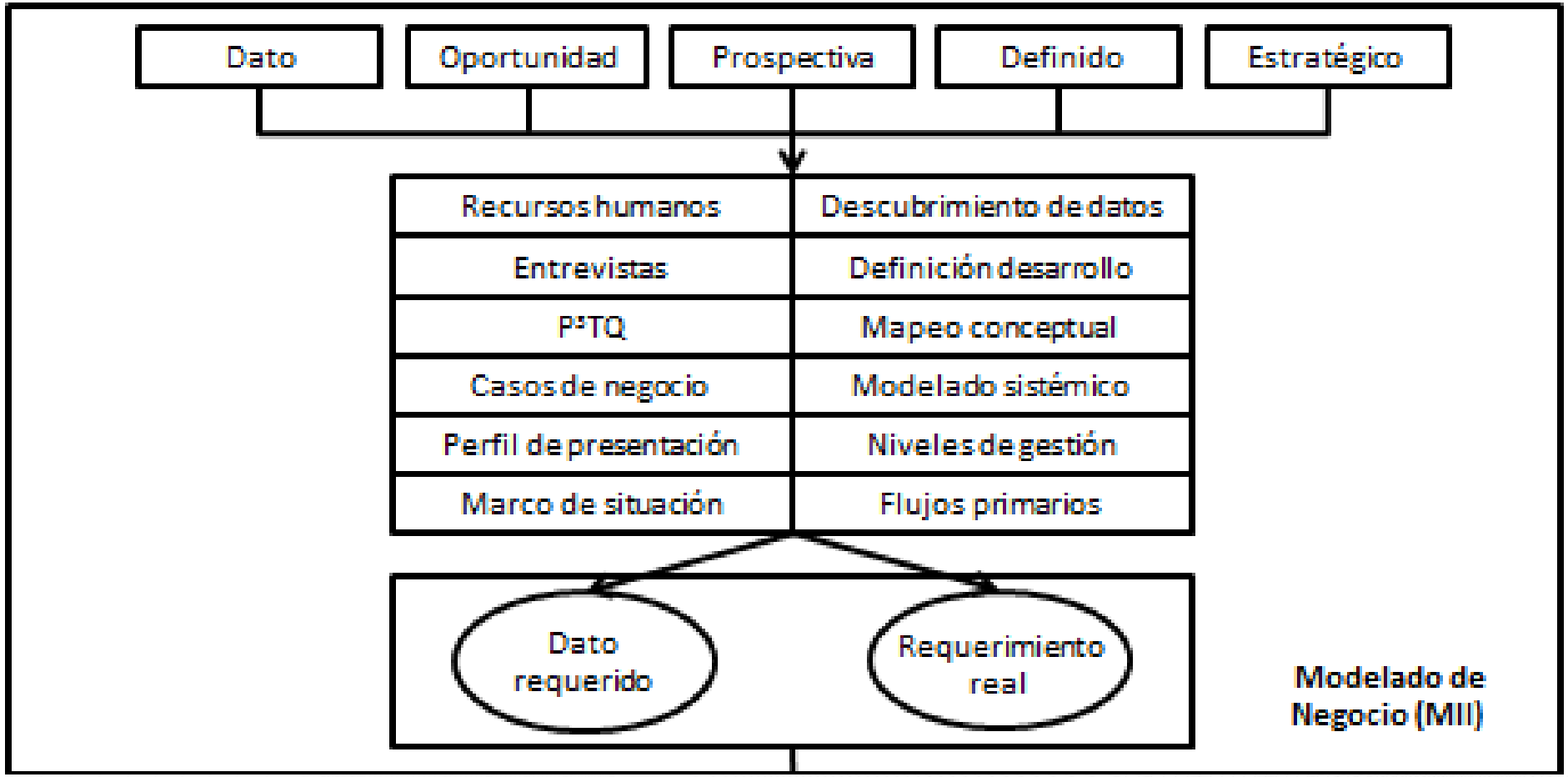
# **Metodología SEMMA (***Sample, Explore, Modify, Model and Assess***)**

# Metodología SEMMA



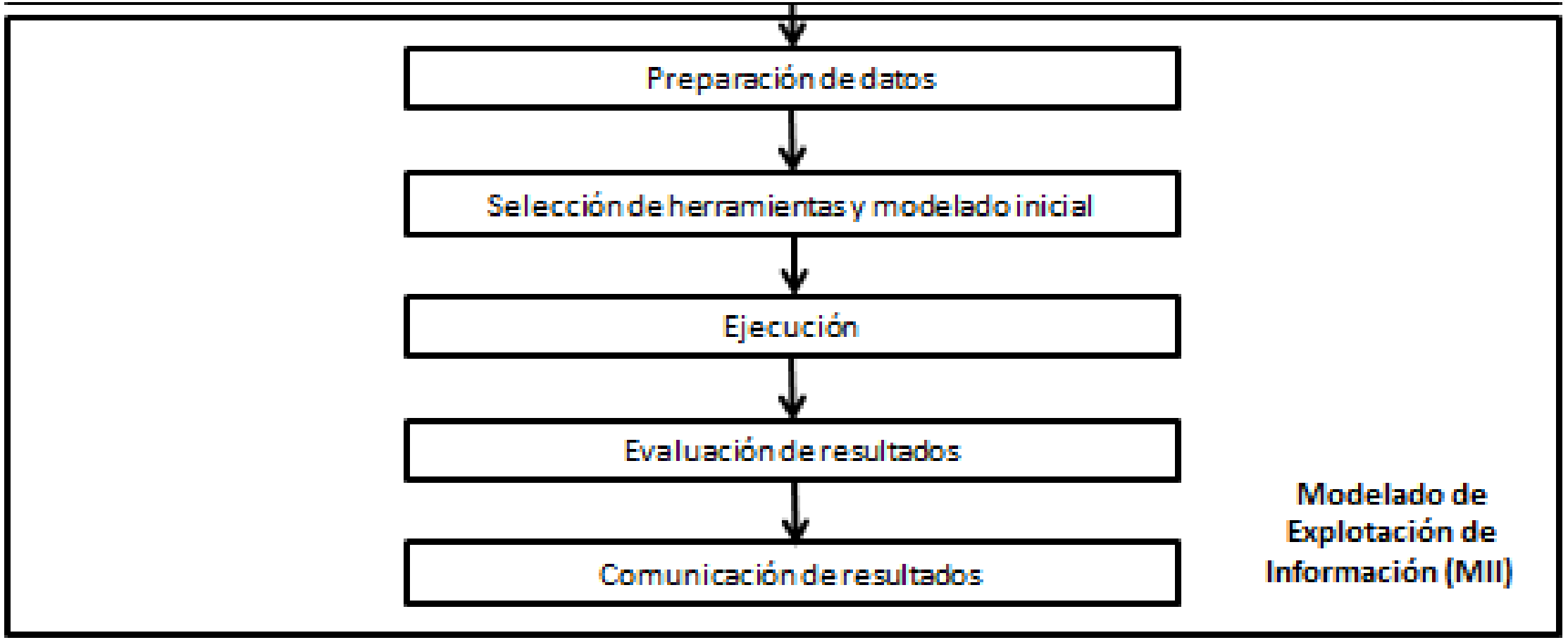
# Metodología Catalyst ó P3TQ

(Producto – Lugar – Precio – Tiempo – Cantidad)

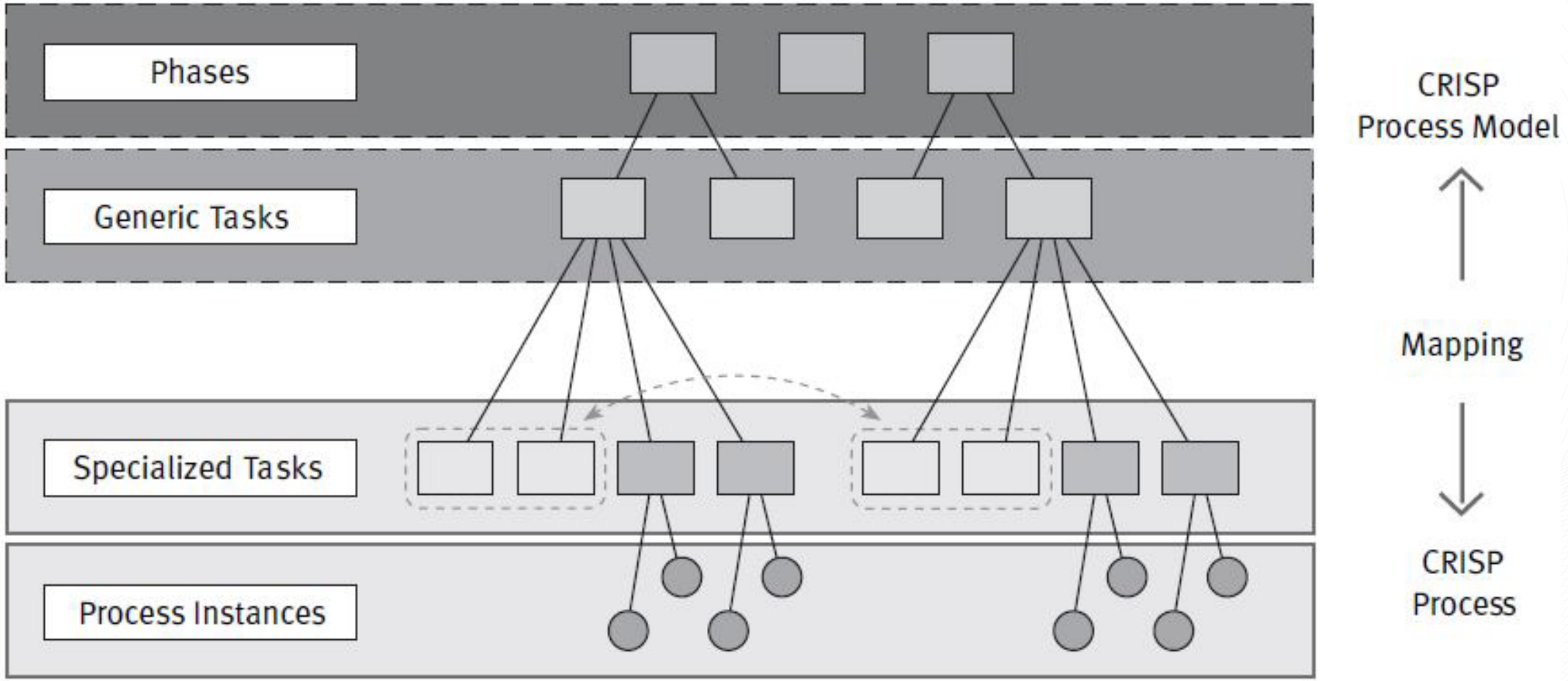


# Metodología Catalyst ó P3TQ

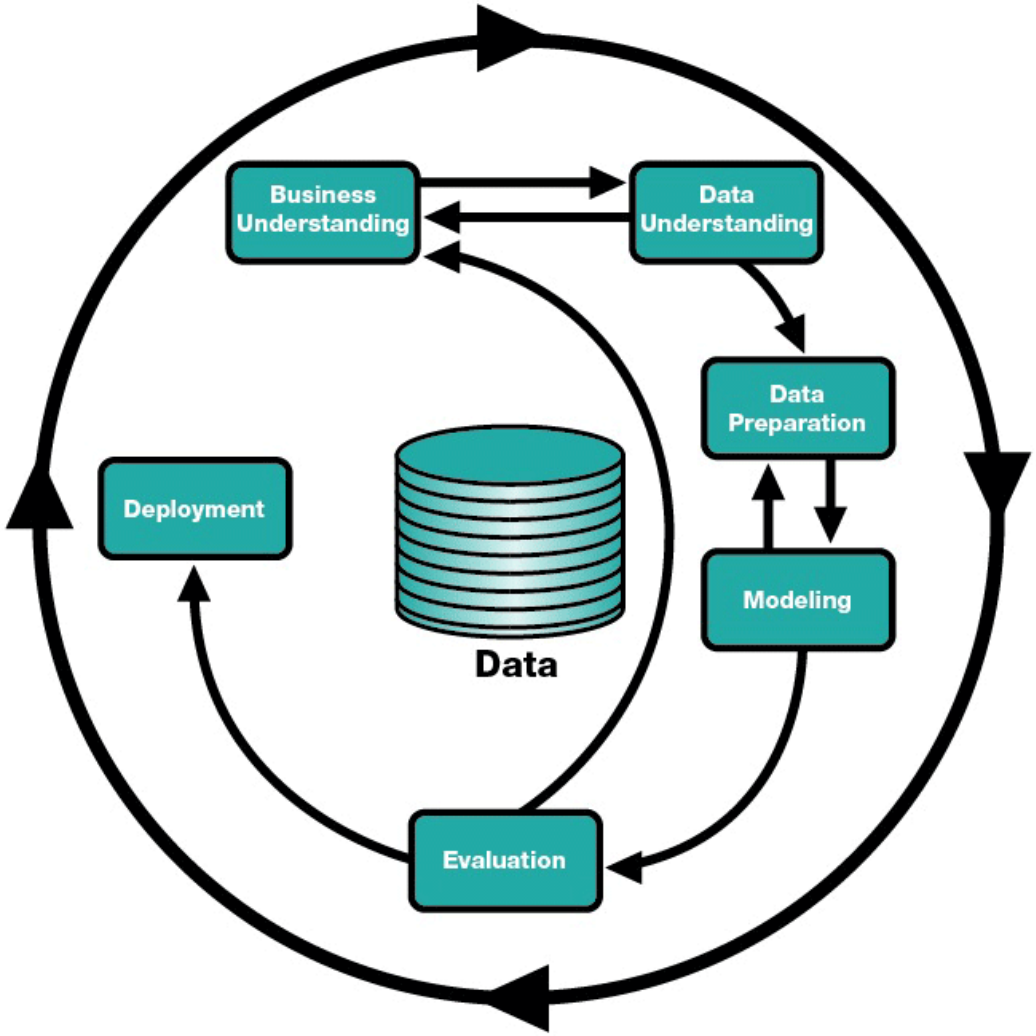
(Producto – Lugar – Precio – Tiempo – Cantidad)



# **Metodología CRISP-DM (***Cross Industry Standard Process for Data Mining***)**



# **Metodología CRISP-DM (***Cross Industry Standard Process for Data Mining***)**



# Metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*)

| FASE | TAREAS GENERICAS | TAREAS ESPECIFICAS |
|---|---|---|
| Comprensión del negocio | Determinar los objetivos del negocio | Background |
| | | Objetivos del negocio |
| | | Criterios de éxito del negocio |
| | Evaluar la situación | Inventarios de recursos |
| | | Requisitos, supuestos y requerimientos |
| | | Riesgos y contingencias |
| | | Terminología |
| | | Costos y beneficios |
| | Determinar objetivos del proyecto de Explotación de Información | Las metas del Proyecto de Explotación de Información |
| | | Criterios de éxito del Proyecto de Explotación de Información |
| | Realizar el Plan del Proyecto | Plan de proyecto |
| | | Valoración inicial de herramientas |
| Comprensión de los datos | Recolectar los datos Iniciales | Reporte de recolección de datos iniciales |
| | Descubrir datos | Reporte de descripción de los datos |
| | Explorar los datos | Reporte de exploración de datos |
| | Verificar la calidad de datos | Reporte de calidad de datos |

# Metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*)

| FASE | TAREAS GENERICAS | TAREAS ESPECIFICAS |
|---|---|---|
| Preparación de los datos | Caracterizar el conjunto de datos | Conjunto de Datos |
| | | Descripción del Conjunto de Datos |
| | Seleccionar los datos | Inclusión / exclusión de datos |
| | Limpiar los datos | Reporte de calidad de datos limpios |
| | Estructurar los datos | Derivación de atributos |
| | | Generación de registros |
| | Integrar los datos | Unificación de datos |
| | Caracterizar el formato de los datos | Reporte de calidad de los datos |
| Modelado | Seleccionar una técnica de modelado | La técnica modelada |
| | | Supuestos del modelo |
| | Generar el plan de pruebas | Plan de pruebas |
| | Construir el modelo | Configuración de parámetros |
| | | Modelo |
| | | Descripción del modelo |
| | Evaluar el modelo | Evaluar el modelo |
| | | Revisación de la configuración de parámetros |

# **Metodología CRISP-DM (***Cross Industry Standard Process for Data Mining***)**

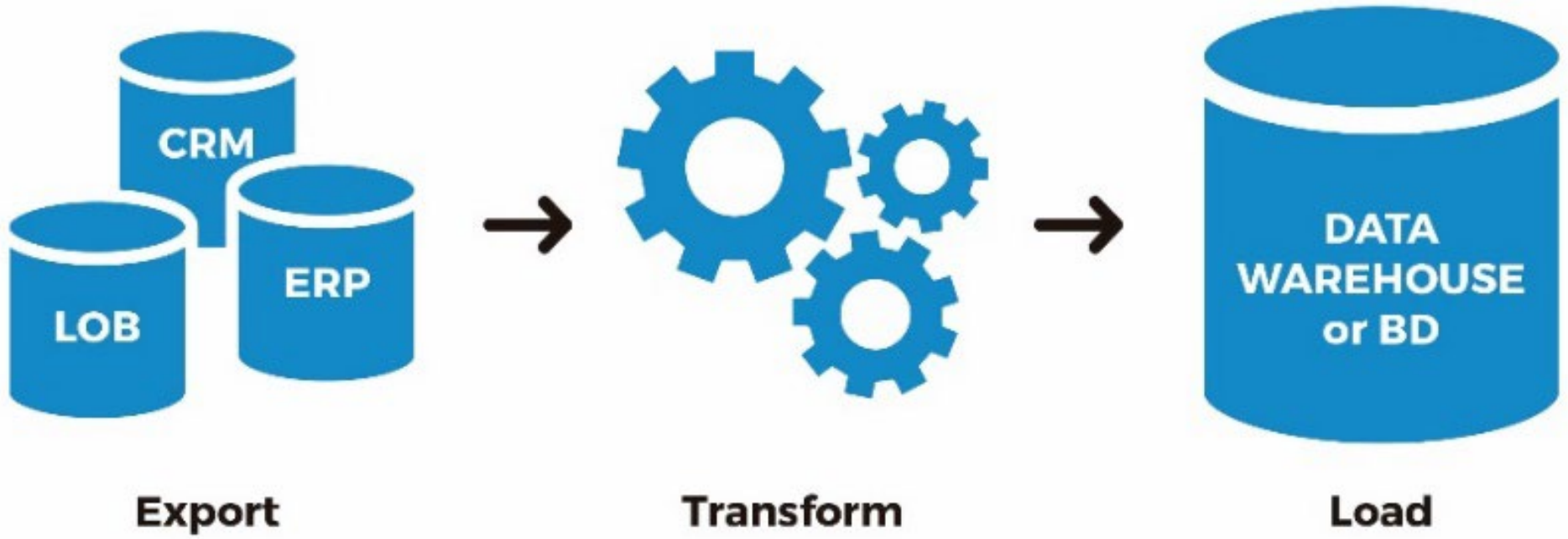
| FASE | TAREAS GENERICAS | TAREAS ESPECIFICAS |
|---|---|---|
| Evaluación | Evaluar Resultado | Valoración de resultados mineros con respecto al éxito del negocio |
| | | Modelos aprobados |
| | Revisar | Revisión del proceso |
| | Determinar próximos pasos | Listar posibles acciones |
| Implementación | Realizar el plan de implementación | Plan de Implementación |
| | Realizar el plan de monitoreo y mantenimiento | Plan de monitoreo y mantenimiento |
| | Realizar el informe final | Informe final |
| | | Presentación Final |
| | Realizar la revisión del proyecto | Documentación de la experiencia |

# Comparación entre Metodologías

| Fases | KDD | SEMMA | CRISP-DM | P3TQ |
|---|---|---|---|---|
| Análisis y comprensión del negocio | • Comprensión del dominio de la aplicación | | • Comprensión del negocio | • Modelado de negocio |
| Selección y preparación de los datos | • Crear el conjunto de datos<br>• Limpieza y pre-procesamiento de los datos<br>• Reducción y proyección de los datos | • Muestreo<br>• Comprensión<br>• Modificación | • Entendimiento de los datos<br>• Preparación de los datos | • Preparación de los datos |
| Modelado | • Determinar la tarea de minería<br>• Determinar el algoritmo de minería<br>• Minería de datos | • Modelado | • Modelado | • Selección de herramientas de modelado inicial |
| Evaluación | • Interpretación | • Valoración | • Evaluación | • Refinamiento del modelo |
| Implementación | • Utilización del nuevo conocimiento | | • Despliegue | • Comunicación |

# Proceso ETL

# Proceso ETL

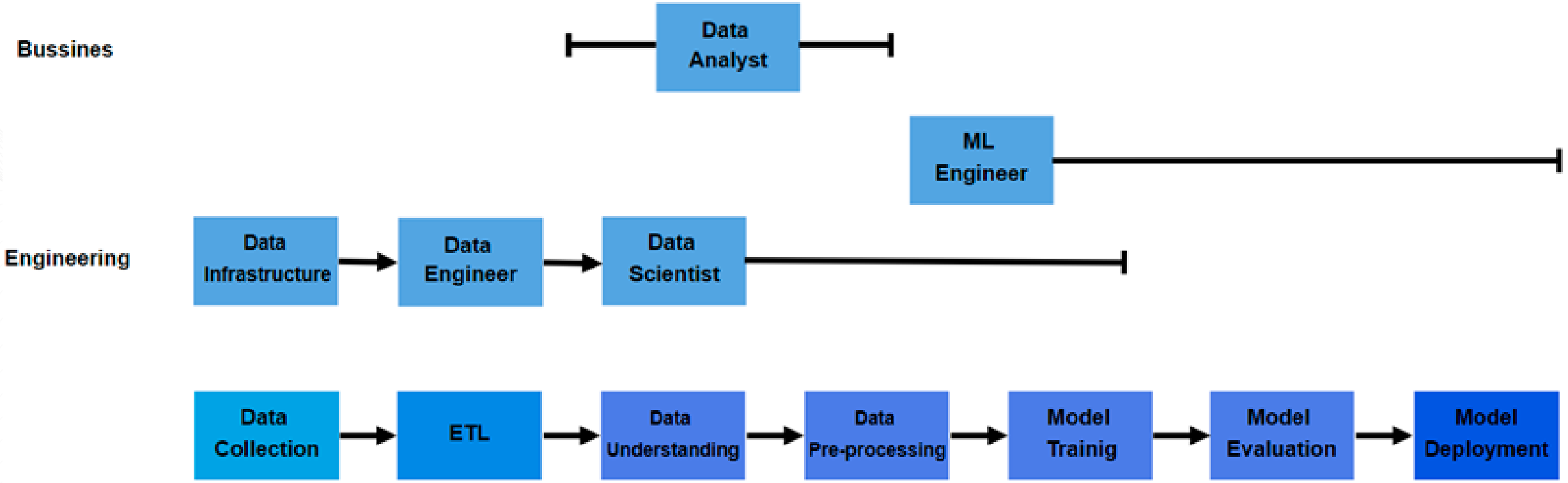


# Roles en la Industria

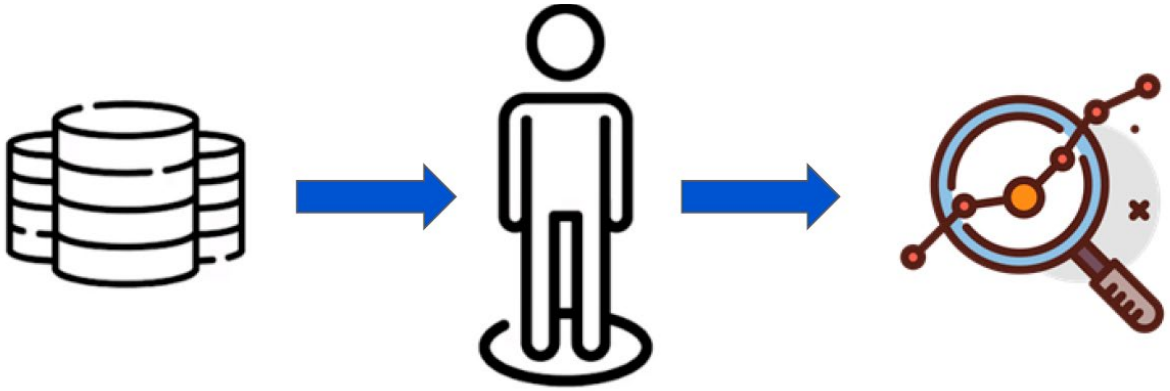
# Roles en la Industria

- Data Science
- Data Analyst
- Data Engineer
- Machine Learning Engineer

# Flujo de Trabajo



# Analista de datos




Identify business needs

Extract data

Visualize and report results

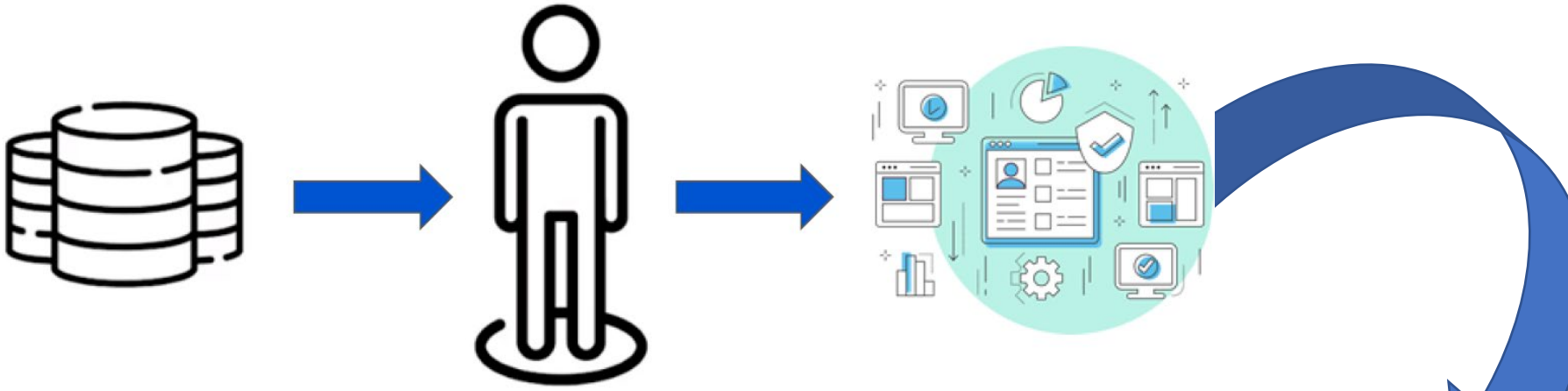
# Analista de datos - Tools

- Software de visualización como Power BI y Tableau
- Lenguajes de programación como Python y R
- Consultas a bases de datos SQL
- Excel

- Estadística descriptiva
- Probabilidades




# Científico de datos



Data

Data Scientist

ML Service /insights

Make decisions based on data.


Incorporate data to software products

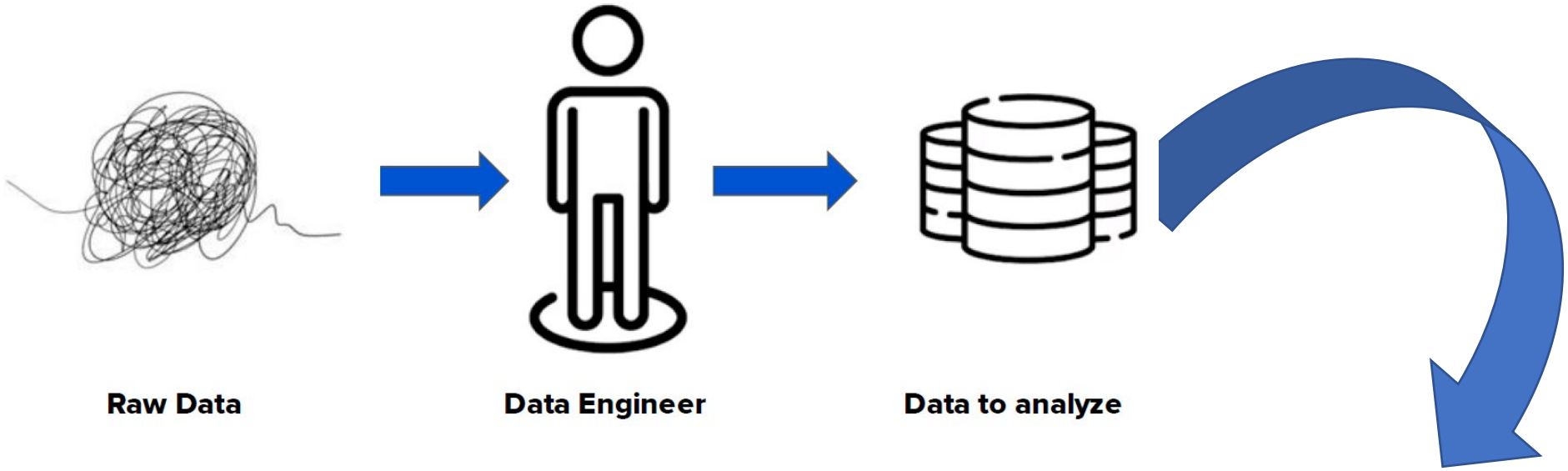
# Científico de datos - tools

- Extraer, limpiar, transformar datos.
- Diseñar, entrenar y evaluar modelos de aprendizaje automático.
- Crear productos de software basados en datos e IA.

- Lenguajes de programación como Python y R
- Consultas a bases de datos SQLPaquetes para manipulación y análisis de datos como Pandas
- Paquetes para ML y DL como como scikit learn y pytorch


- Álgebra lineal
- Probabilidad y estadística avanzadas
- Cálculo

# Ingeniero de datos



Raw Data

Data Engineer

Data to analyze

**Works for the team to have data to be analyzed**

**Build ETLs**

# Ingeniero de datos - tools

- Lenguajes de programación como python.
- Conocimientos avanzados en bases de datos.
- Tecnologías para la gestión de Big Data como Spark , y Delta Lake



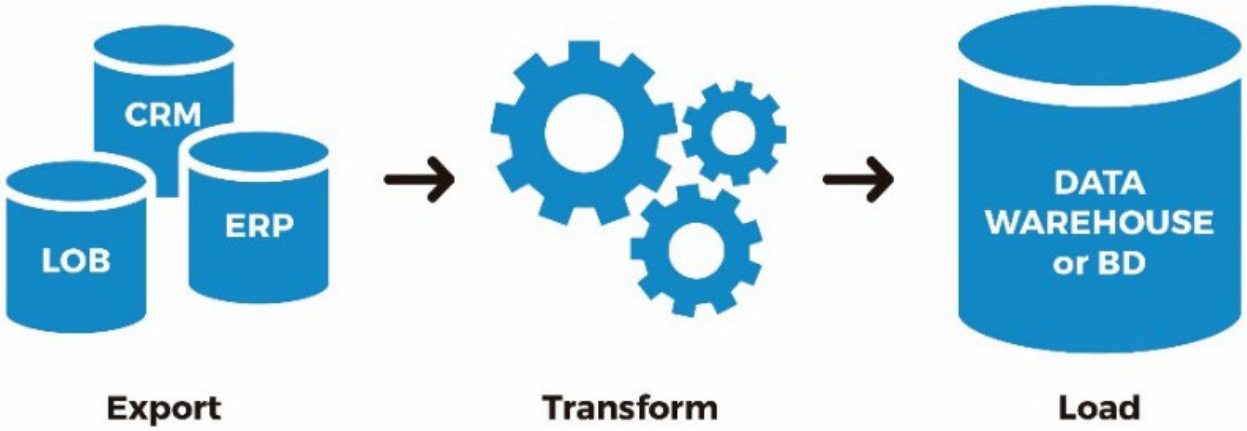


- Cloud platforms.
- Docker containers.
- Kubernetes




- Estadística descriptiva

# Procesos de Big Data - Analitics




## Preparación de los Datos

1. Integración de los datos
2. Eliminar variables irrelevantes y redundantes
3. Descripción estadística de los datos
4. Limpieza de datos
5. Transformación de tipo de datos según el método

# 1. Integración de datos


+


+

# 2. Eliminar variables irrelevantes y redundantes

- Nombre
- Teléfono
- Dirección
- Documento de Identificación

Variables categóricas:
– Edad={niño, adolescente, adulto}
– Mayor de edad={si, no}

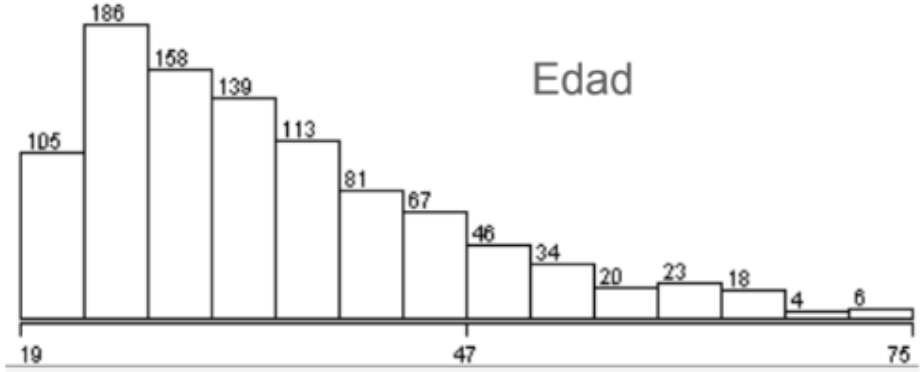
Variables numéricas:
– Edad
– Año de nacimiento

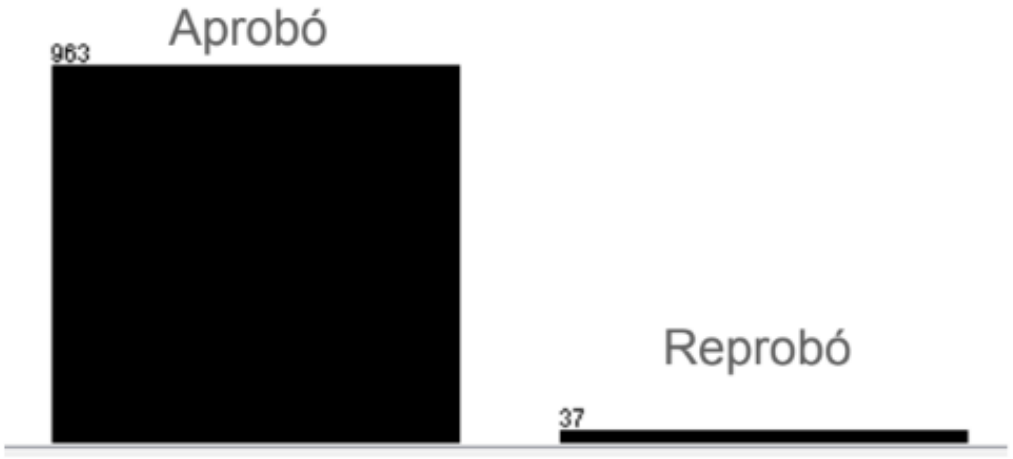
Variables numéricas/categóricas:
– Edad
– Mayor de edad={si, no}

# 3. Descripción estadística de los datos

Edad

| Statistic | Value |
|---|---|
| Minimum | 19 |
| Maximum | 75 |
| Mean | 35.546 |
| StdDev | 11.375 |



# 4. Limpieza de datos

Datos atípicos (outliers)
Datos faltantes (nulos)

| Id | Nombre | Estrato | Sexo | Enfermedad | Colegio_U | Activo_Web | Asistencia | Entregas_Completas | Trabaja | Examen_Final |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ana Perez | 3 | F | SI | SI | BAJA | 0,1 | 0,45 | NO | APROBADO |
| 2 | Mauricio Rios | 2 | M | NO | NO | MEDIA | 0,45 | 0,6 | NO | APROBADO |
| 3 | Samuel Ochoa | 4 | M | NO | NO | ALTA | 0,5 | 0,75 | SI | DESAPROBADO |
| 4 | Emilia Oviedo | 3 | F | NO | NO | ALTA | 0 | 0,5 | SI | DESAPROBADO |
| 5 | Carmen Reyes | 4 | F | SI | NO | MEDIA | 0,65 | 0,85 | NO | APROBADO |
| 6 | Alberto Arenas | 4 | M | NO | NO | BAJA | 0,1 | 0 | NO | DESAPROBADO |
| 7 | Maria Betancur | 3 | M | NO | NO | MEDIA | 0,2 | 0,9 | NO | APROBADO |
| 8 | Manuel Regino | 2 | M | NO | NO | MEDIA | 0,3 | 0,8 | SI | DESAPROBADO |
| 9 | Sara Merino | | F | NO | SI | BAJA | 0,35 | 0,7 | | APROBADO |
| 10 | Pepe Corrales | 2 | M | SI | NO | BAJA | 0,75 | 0,5 | SI | DESAPROBADO |
| 11 | Raul Poveda | 4 | M | NO | SI | ALTA | 0,7 | 0,6 | NO | APROBADO |
| 12 | Cristina Carrillo | 3 | F | NO | NO | MEDIA | 0,9 | 0,8 | NO | APROBADO |
| 13 | Olga Arias | 2 | F | NO | NO | ALTA | 0,2 | 0,25 | NO | DESAPROBADO |
| 14 | Yaned Cardona | 4 | F | SI | NO | ALTA | 0,2 | 0,2 | NO | DESAPROBADO |
| 15 | Paula Quintero | 3 | F | NO | NO | BAJA | 0,9 | 0,8 | NO | APROBADO |
| 16 | Jeronimo Martinez | 2 | M | NO | NO | MEDIA | 1 | 1 | NO | APROBADO |

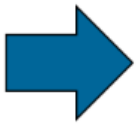
- Eliminar los registros
- Eliminar la variable – columna
- Completar los valores faltantes
- Imputar: Completar los campos con la media o la moda

# 5. Transformación de tipo de datos según el método

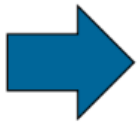
## Conversión de número a categorías

**Discretización**

| Edad Numérica |
|---|
| 12 |
| 17 |
| 24 |
| 15 |
| 7 |



0 – 12 -> Niño
13 – 17 -> Adolescente
18 en adelante -> Adulto

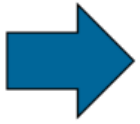


| Edad Categórica |
|---|
| Niño |
| Adolescente |
| Adulto |
| Adolescente |
| Niño |

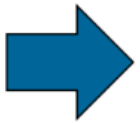
## Conversión de categorías a números

**Codificador**

| Edad Categórica |
|---|
| Niño |
| Adolescente |
| Adulto |
| Adolescente |
| Niño |



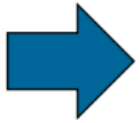
Niño -> 1
Adolescente -> 2
Adulto -> 3



| Edad Numérica |
|---|
| 1 |
| 2 |
| 3 |
| 2 |
| 1 |

**Creción de Dummies**

| Ciudad |
|---|
| Medellín |
| Bogotá |
| Cali |
| Cali |
| Cartagena |



| Bogotá | Cartagena | Medellín | Cali |
|---|---|---|---|
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 |

# G R A C I A S

**Presentó**: Alvaro Pérez Niño
Instructor Técnico
**Correo**: aperezn@misena.edu.co
http://centrodeserviciosygestionempresarial.blogspot.com/