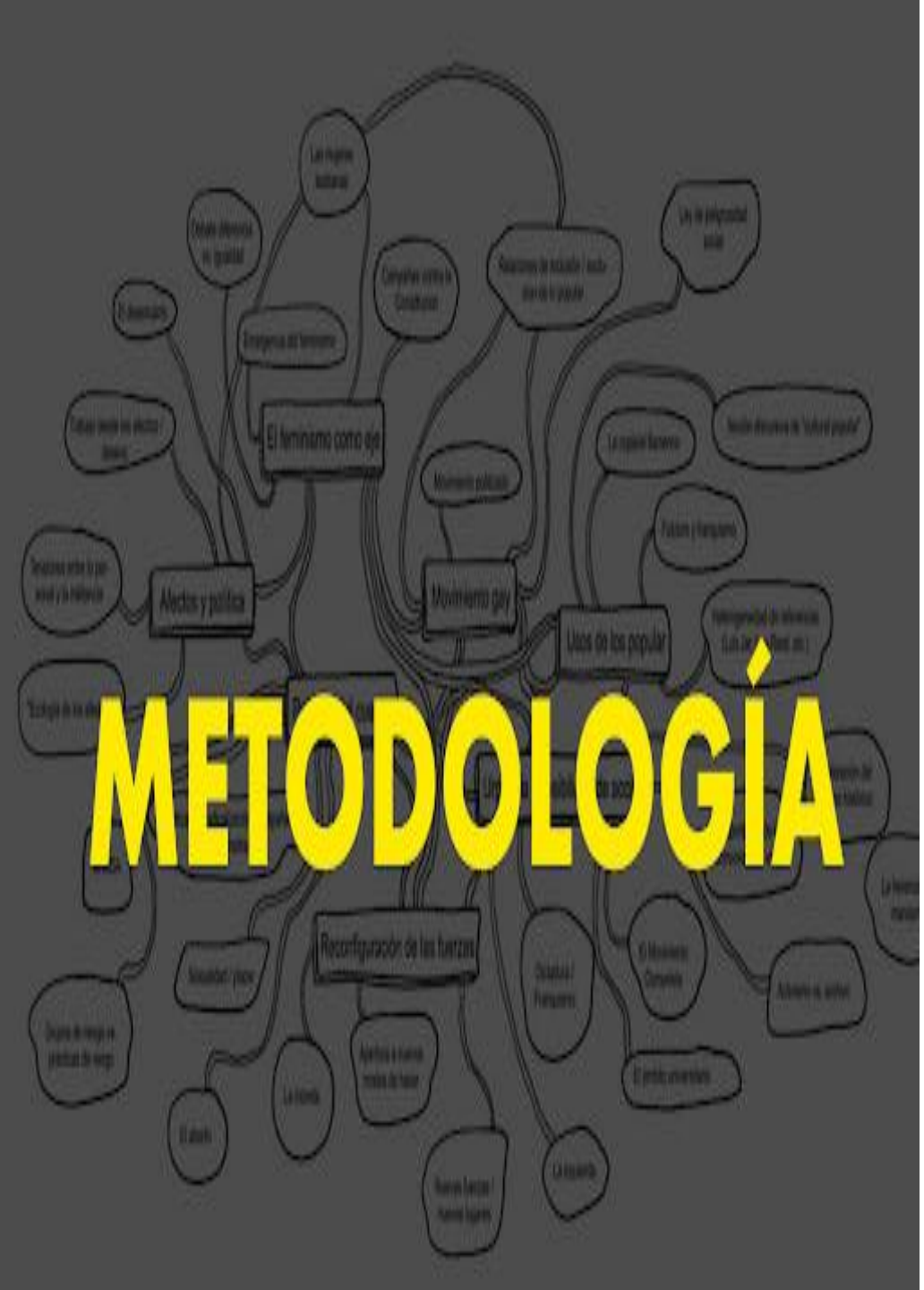
Semilleros
by DSRP

Metodologías



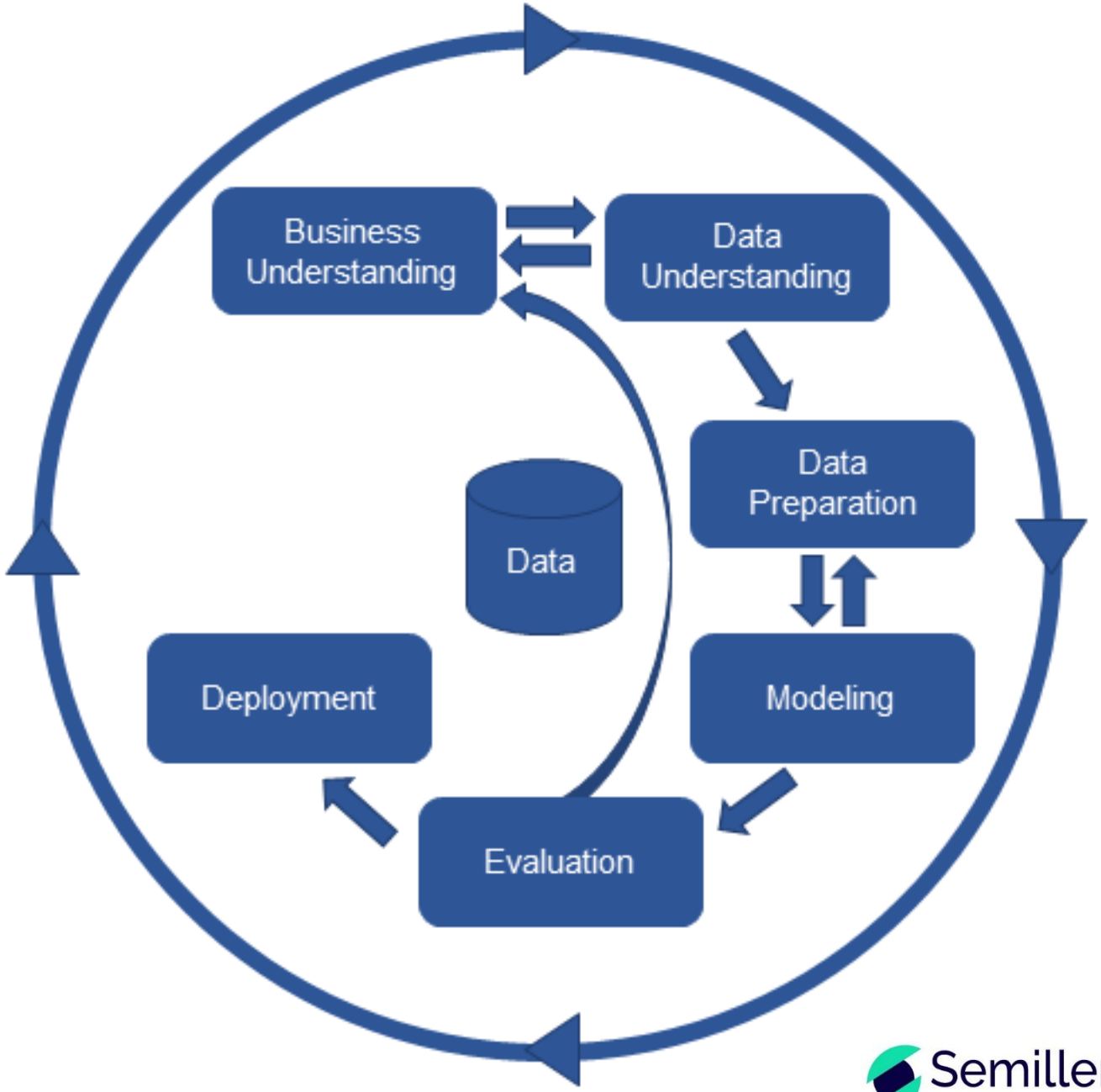
METODOLOGÍA CRISP - DM

Semilleros by DSRP

METODOLOGÍA
CRISP - DM

Cross-Industry Standard Process for Data Mining

(finales 90’s)



Semilleros by DSRP

Fase 1. Business Understanding: Definición de necesidades del cliente (Comprensión del negocio)

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase 2. Data Understanding: Estudio y comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Semilleros by DSRP

Fase 3. Data Preparation. Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Fase 4. Modeling. Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Semilleros by DSRP

Fase 5. Evaluation. Evaluación (obtención de resultados)

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.
Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase 6. Deployment. Despliegue (puesta en producción)

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

Fases y Actividades

Comprensión del Negocio

- **Determinar los Objetivos del Negocio**
  - ✓ *Antecedentes*
  - ✓ *Objetivos del Negocio*
  - ✓ *Criterio de Éxito*
- **Evaluar la situación**
  - ✓ *Inventario de requerimientos de Recursos, Hipótesis y Limitaciones*
  - ✓ *Riesgos y Contingencias*
  - ✓ *Terminología*
  - ✓ *Costos y Beneficios*
- **Determinar el objetivo de Minería de Datos**
  - ✓ *Objetivos de Minería de Datos*
  - ✓ *Criterio de Éxito de Minería de Datos*
- **Desarrollar el Plan de Proyecto**
  - ✓ *Plan de proyecto*
  - ✓ *Evaluación inicial de*

Comprensión de Datos

- **Obtener los datos iniciales**
  - ✓ *Reporte de la obtención de los datos iniciales*
- **Describir los Datos**
  - ✓ *Reporte con la descripción de los datos*
- **Explorar de Datos**
  - ✓ *Reporte de la Exploración de Datos*
- **Verificar de la calidad de los Datos**
  - ✓ *Reporte de la calidad de los datos*

Preparación de Datos

- *Conjunto de Datos*
- *Descripción de los Datos*
- **Seleccionar los Datos**
  - ✓ *Justificación de la inclusión / Exclusión*
- **Limpiar Datos**
  - ✓ *Reporte de Limpieza de Datos*
- **Construir Datos**
  - ✓ *Atributos Derivados*
  - ✓ *Registros Generados*
- **Integrar Datos**
  - ✓ *Datos Combinados*
- **Dar formato a los Datos**
  - ✓ *Datos Formateados*

Modelamiento

- **Seleccionar Técnica de Modelamiento**
  - ✓ *Técnica de Modelamiento*
  - ✓ *Modelamiento*
  - ✓ *Hipótesis*
- **Generar el Diseño de Prueba**
  - ✓ *Diseño de Prueba*
- **Construir el Modelo**
  - ✓ *Configuración de los parámetros del Modelo*
  - ✓ *Descripción del Modelo*
- **Evaluar el Modelo**
  - ✓ *Evaluación del Modelo*
  - ✓ *Revisión de la configuración de los parámetros del modelo*

Evaluación

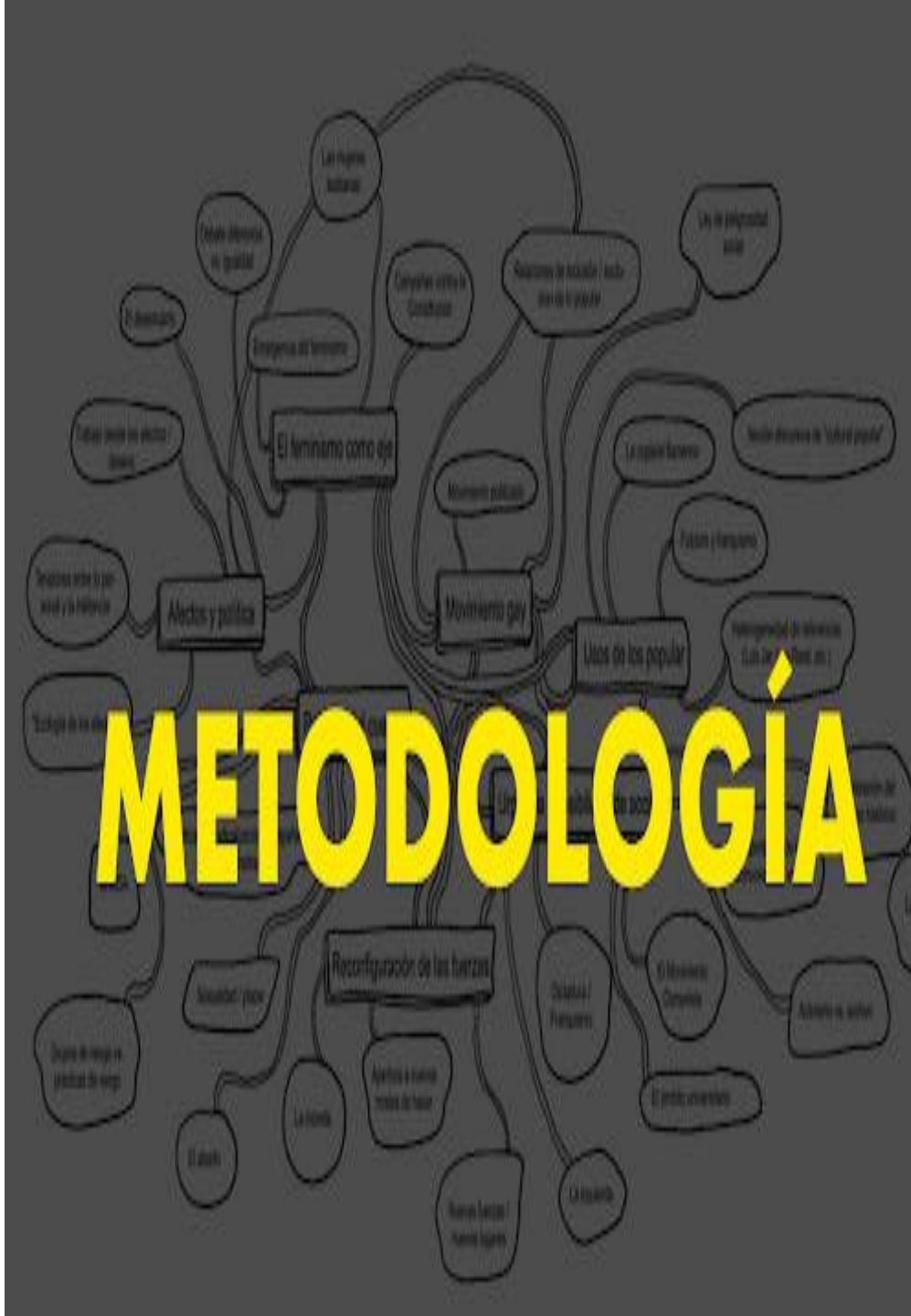
- **Evaluar Resultados**
  - ✓ *Hipótesis de Minería de Datos*
  - ✓ *Resultados*
  - ✓ *Criterio de éxito del negocio*
  - ✓ *Modelos aprobados*
- **Revisar el Proceso**
  - ✓ *Revisión del Proceso*
- **Determinar los siguientes pasos**
  - ✓ *Lista de Posibles Acciones*
  - ✓ *Decisión*

Despliegue

- **Desplegar el Plan**
  - ✓ *Plan de Despliegue*
- **Monitorear y Mantener**
  - ✓ *Plan de monitoreo y Mantenimiento*
- **Desarrollar el reporte final**
  - ✓ *Reporte Final*
  - ✓ *Presentación Final*
- **Revisión del Proyecto**
  - ✓ *Documentación de las experiencias*

Lesson #2

Metodología KDD - Process



Semilleros by DSRP

KDD Process
(Knowledge Discovery in Databases)

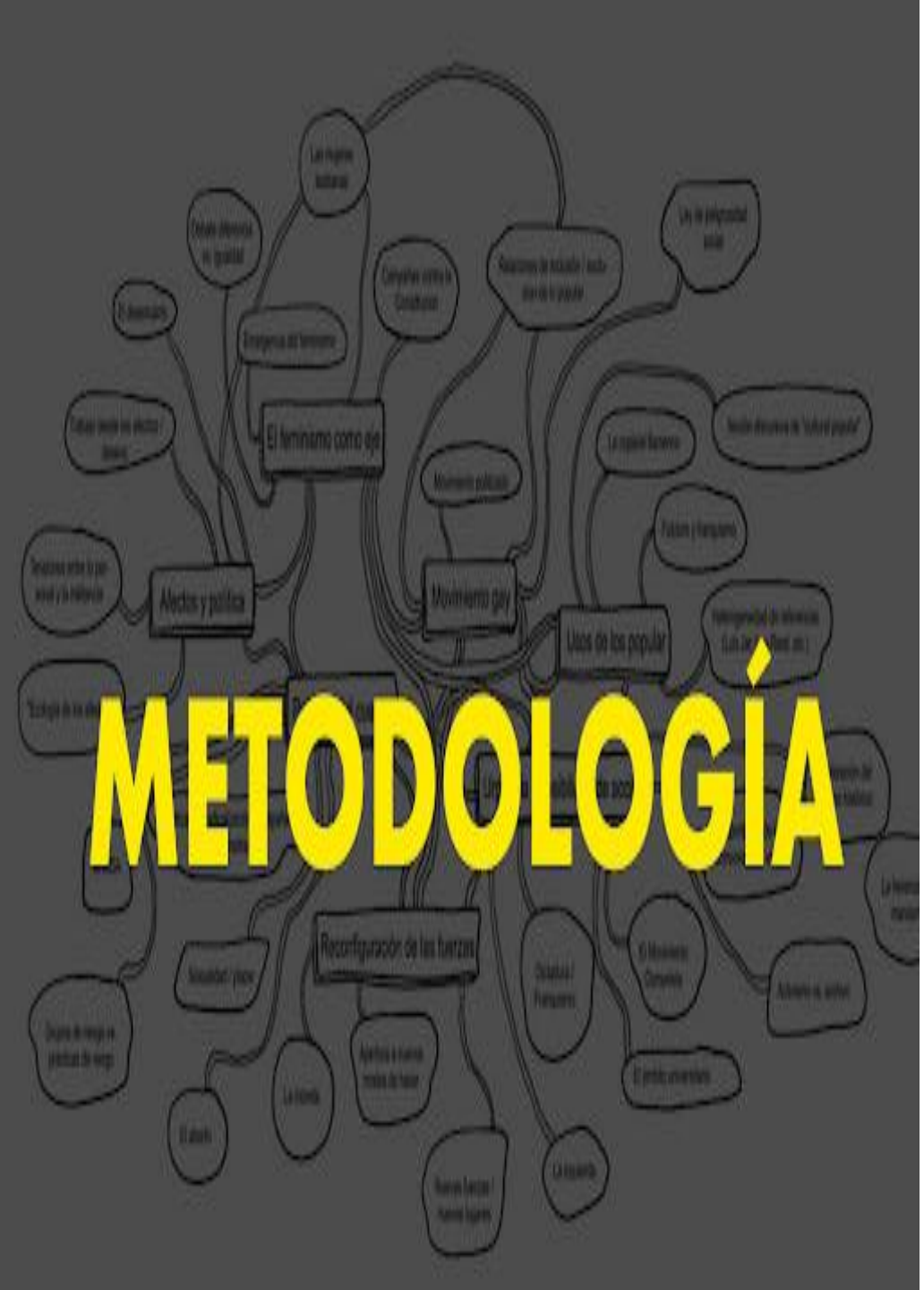
El término KDD es una forma de referirnos a la obtención del conocimiento a partir de una colección de datos.

Semilleros by DSRP

Fases de la Metodología

Selección de datos	Se recolectan los datos necesarios, y se pasan al formato requerido.
Pre-Procesamiento	Corresponde a la limpieza y normalización de los datos obtenidos.
Transformación	Incluye la división de los datos de prueba y de entrenamiento, y la implementación de los algoritmos a utilizar.
Minería de datos	Es la fase de entrenamiento y aprendizaje de los modelos
Interpretación y Evaluación	Los datos obtenidos se verifican y se compara el rendimiento de los modelos.

Semilleros by DSRP



METODOLOGÍA SEMMA

Semilleros by DSRP

Fases y actividades SEMMA

Entrada de datos,
Ejemplos,
Partición de datos

Exploración distribuida,
Múltiples particiones,
Intuición, Asociación,
Selección de Variables

Transformación de variables, Filtros a los datos fuera de rango, Agrupamiento, ruido

Regresión,
Arboles,
Redes neuronales
Etc.

Evaluación,
Medidas,
Reportes

Semilleros by DSRP

Comparación entre KDD, SEMMA y CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	xxxxx	Conocimiento del negocio
Selección	muestra	Conocimiento de los datos
Preprocesamiento	exploración	
Transformación	Modificación	
Minería de datos	Modelo	
interpretación / evaluación	evaluación	
Post KDD	xxxxx	

Semilleros by DSRP