# From Business Problems to Data Mining Tasks

Descomponer el problema en partes pequeñas para las cuales ya existen herramientas y métodos.

1. Clasificación: Clasifica una entidad en grupos mutuamente excluyentes.

Estimación de probabilidad (scoring): Clasifica la probabilidad de que una entidad pertenezca a un grupo.

Predice si algo va a pasar o no.

1. Regresión: Predice cuánto va a pasar.
2. Similarity matching: Busca entidades con características similares.
3. Clustering: Agrupar individuos con base en sus similitudes.
4. Co-occurrence grouping: Busca qué entidades suelen estar juntas.
5. Profiling: Caracterizar un comportamiento típico de la entidad.
6. Link prediction: Predice conexiones que deberían existir basados en los datos.
7. Data reduction: Reducir el dataset para obtener información más específica. Suele haber pérdida de información. Toca evaluar las ganancias.
8. Causal modeling: Qué eventos o acciones influencian a otras. Relación causal. Siempre deben haber supuestos que deben cumplirse.

# Supervised Versus Unsupervised Methods

Supervisado se da un objetivo, en el otro no. El no supervisado saca sus propias conclusiones.

Classification, regression, and causal modeling generally are solved with supervised methods. Similarity matching, link prediction, and data reduction could be either. Clustering, co-occurrence grouping, and profiling generally are unsupervised.

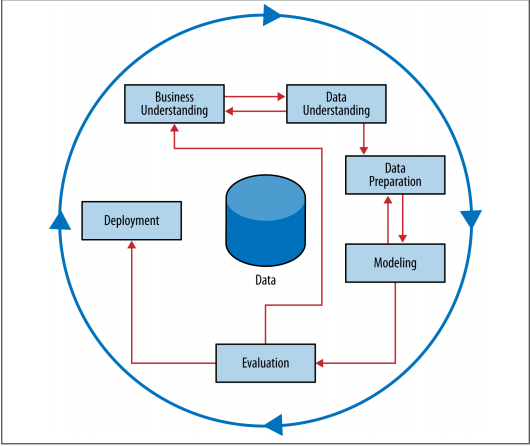
# Data Mining and Its Results

DM: Usar los datos para encontrar patrones y crear modelos.

IR: Usar el modelo obtenido arriba para llenar campos vacíos.

# The Data Mining Process

CRISP-DM <- Base de ASUM-DM vista con Tamura.



## Business Understanding

Entender el problema. Formularlo bien. Estructurarlo de manera que se puedan usar métodos existentes.

## Data Understanding

Ver los datos. Determinar si sirven. Determinar si se necesitan más. Determinar cómo se relacionan con los métodos existentes.

## Data Preparation

Lo de la clase de Norha más revisar leaks. Variables que sirven para predecir pero no se pueden obtener en tiempo real (Leaks).

## Modeling

Lo de la clase de Julio.

## Evaluation

Evaluar al máximo los modelos para asegurarse de que no serán catastróficos. Asegurarse que harán más bien que mal. Tener en cuenta que al poner en práctica los modelos el mundo puede cambiar y el modelo podría quedar obsoleto. (Por aquí puede haber algo de Autoadaptativos)

## Deployment

Despliegue 😐

# Implications for Managing the Data Science Team

No es lo mismo que un proyecto de software. Va más orientado a la investigación, por lo tanto es un ciclo de retroalimentación.

Las habilidades de software son diferentes a las habilidades analíticas.

# Other Analytics Techniques and Technologies

## Statistics

Lo de la clase de Julio

## Database Querying

Lo que yo hago

## Data Warehousing

Obtienen datos de varias bases de datos. DWs are central repositories of integrated data from one or more disparate sources.

## Regression Analysis

Lo de la clase de Julio

## Machine Learning and Data Mining

IA -> Machine Learning -> KDD (Knowledge Discovery and Data Mining)

Técnica y algoritmos para reconocimiento de patrones.

## Answering Business Questions with These Techniques

¿Quiénes ganan más? SQL

¿Hay diferencias entre los que más ganan y los que menos ganan para la empresa? Julio

Predecir si un nuevo cliente tendrá las lucas. Clasificación. Data Mining

Qué tanto va a darle a la empresa ese nuevo cliente. Regresión. Data Mining