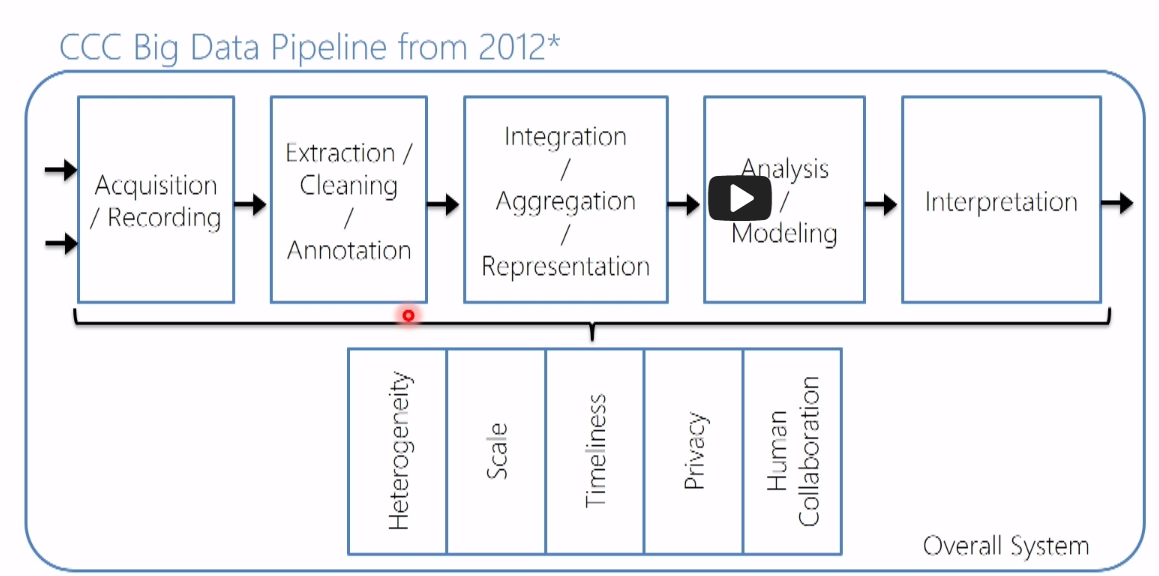
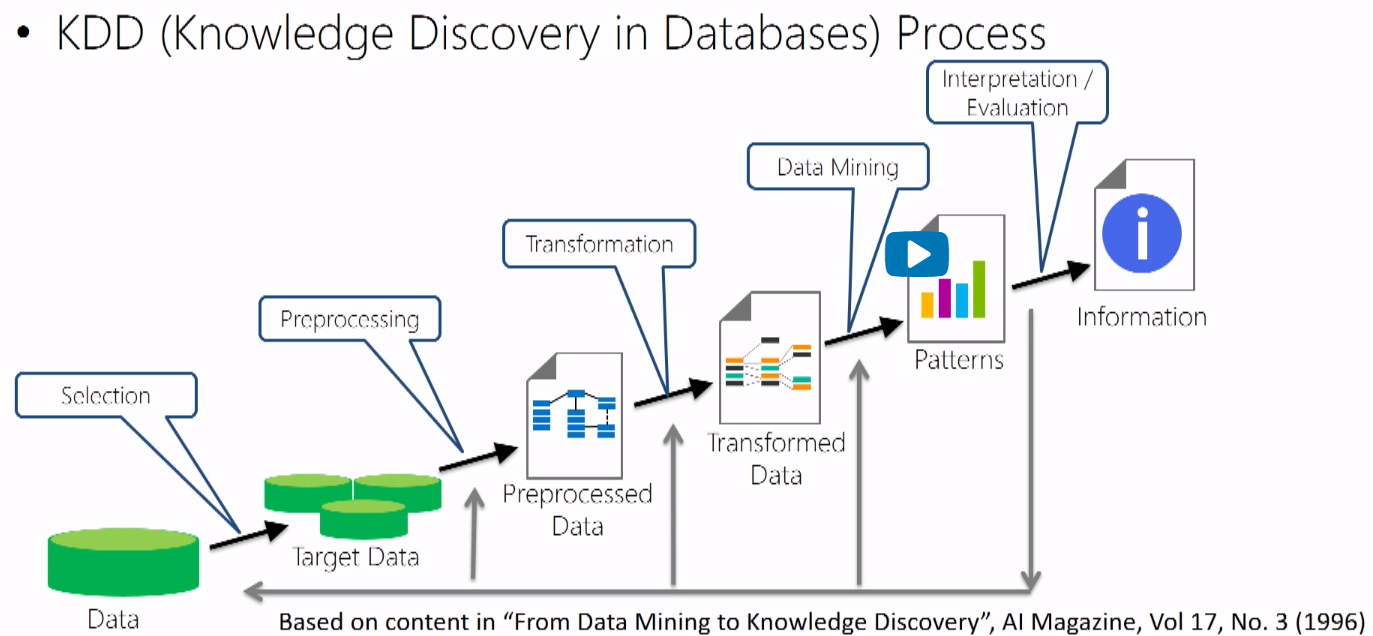
# OVERVIEW:

Notas historicas sobre KDD, CRISP-DM, BigData and DataScience y las relaciones con Data Mining y Machine Learning.



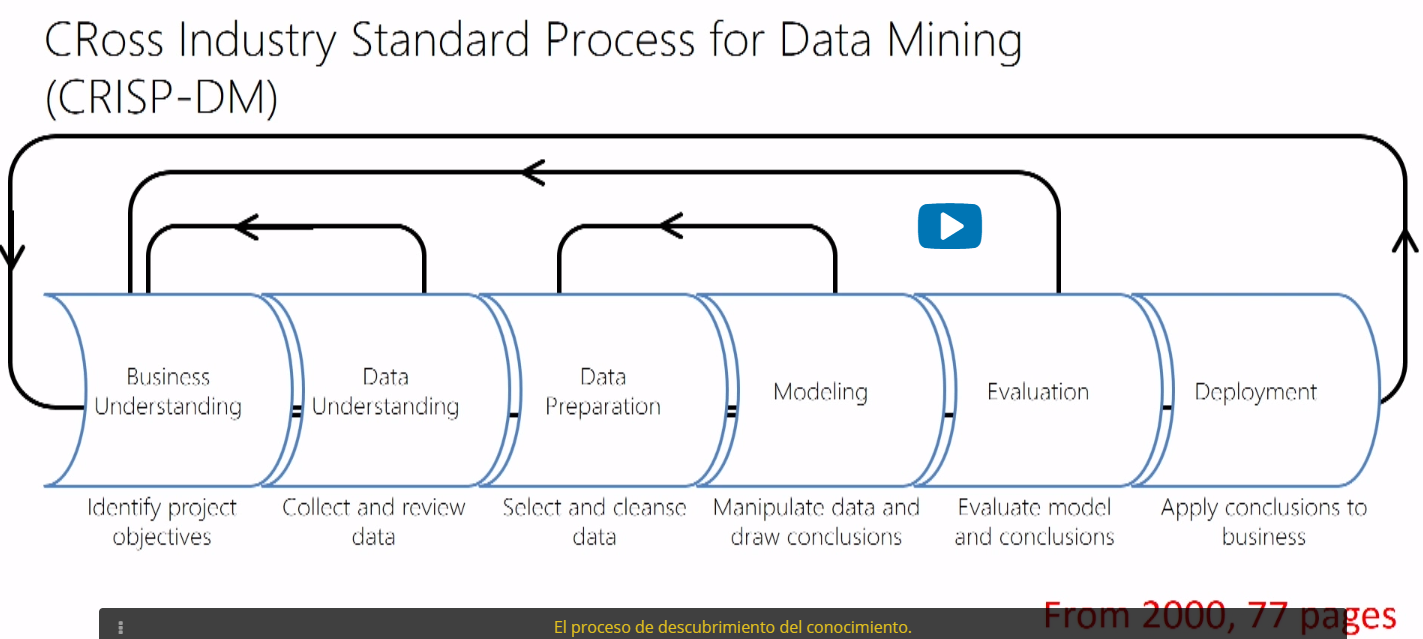
1. **Acquisition / Recording:** Obtención de datos, grabar como se comportan los usuarios en diferentes sitios web o usar script para recopilar los datos correctamente.
2. **Extraction / Cleaning / Annotation:** Se intenta reducir el ruido un poco, deshacerse de los datos que se necesita directamente, o convertir texto en tabla estructurada. Básicamente se esta cambiando los datos de lo que es potencialmente una pila de basura en algo con lo que realmente se puede trabajar.
3. **Integration / Agregation / Representation:** Se configura los datos de una manera directamente conducente a la minería de datos, asi que todos los documentos de texto serán vinculados correctamente a otras tablas estructuradas. Y se tiene otra base de datos central donde toda la información se almacena de forma agradable y ordenada.
4. **Analysis / Modeling:** Donde la gran máquina de aprendizaje y minería de datos se lleva a cabo.
5. **Interpretation:** Es donde al final ves lo que tienes y lo interpretas.

# KDD – Knowledge Discovery in Databases



El modelo es el mismo que el CCC White Paper

# CRISP-DM – Cross Industry Standard Process for Data Mining



# NOTAS HISTORICAS

* Las etapas son básicamente las mismas no importar quien invente o reinvente este proceso de descubrimiento de conocimiento, minería de datos, big data o data science. Puede ser que no siempre necesites todas las etapas.
* Data science es un proceso iterativo
  + Las flechas hacia atrás son parte del proceso.

# EJEMPLO: FALLAS DE ENERGIA EN MANHATTAN

* La demanda máxima de electricidad esta aumentando.
* La infraestructura que soporta NY se remonta a la década de 1880 desde la época de Thomas Edison.
* Las fallas de energía ocurren en NY con bastante frecuencia (basta con hacer estadísticas) y en realidad son caras de reparar.
* El OBJETIVO era descubrir como priorizar los pozos que serian inspeccionados con el fin de reducir de manera optima el número de eventos de registro en el futuro. Los eventos son Incendio, explosión, fumar en un pozo.

## Opportunity Assessment & Business Understanding:

* ¿Qué es lo que realmente se quiere lograr?
* ¿Cuáles son tus limitaciones, cuáles son tus riesgos?
* ¿Cómo vas a evaluar la calidad de tus resultados?
* Los objetivos son concretos, podemos realizar evaluaciones concretas de que hicimos bien las tareas
* Objetivo 1: Predecir eventos de pozo en el año antes de que ocurran, asi especificamos un periodo de tiempo particular y definimos lo que era un evento de pozo.
* Objetivo 2: Realizar un análisis costo beneficio, para implementar políticas de inspección, estas políticas que tienen en cuenta el costo de la inspección y el costo de un incendio, ya que a mas Inspecciones, menos incendios y a menos inspecciones mas incendios. Lo cual determina con que frecuencia se tiene que inspeccionar los pozos.

## Data Understanding & Data Acquisition:

* Los datos que se tenían eran tickets de problemas, los cuales fueron escritos por despachadores de la compañía eléctrica mientras dirigían la acción de solucionar el problema y grababan todo eso mediante eso sucedia.
* Tambien había muchos registros de información sobre los pozos, ubicación, cobertura, información básica.
* Tambien había registros sobre los tipos de cable usados.
* Informacion sobre shocks eléctricos
* Información extra sobre eventos severos
* Reporte de inspecciones
* Datos sobre cobertura de pozo

## Data Cleaning and Transformation:

* Transformar texto libre en información estructurada. (Text Mining)
* Tratar de integrar las tablas (crear identificadores únicos)

## Model Building:

* A menudo los modelos predictivos significan aprendizaje automático o modelización estadística, no necesariamente significa predicción en el futuro.
* Si tu objetivo es responder preguntas de si o no, entonces esto es **CLASIFICACIÓN**
  + **Ejemplo:** ¿Explotará un pozo el próximo año, si o no?
* Si tu objetivo es predecir un valor numérico, esto es **REGRESIÓN**.
  + Si quieres predecir cuantas bicicletas se van a alquilar el próximo año, es un problema de regresión.
* Si quieres agrupar observaciones similares, es un problema de **AGRUPAMIENTO.**
* Si quieres recomendar a alguien un artículo, como un libro, película o un producto basado en las calificaciones de otros clientes. Es un sistema de **RECOMENDACIÓN.**

## Construcción Política:

* ¿Como va a usarse tu modelo para cambiar la política?
  + Para los pozos, ¿cómo deberíamos recomendar el cambio de la política de inspección basada en nuestro modelo?
* **Ejemplo:** Si Starbucks se muda a una nueva ciudad, después de la creación del modelo. ¿Cómo se hace esa optimización donde se encuentra las tiendas?,¿Qué tan grandes deben ser?,¿Dónde deben estar ubicados los almacenes? Es un problema de optimización. Entonces este es un problema de construcción de políticas.
* Construcción de modelo es **PREDICTIVO,** La Construcción de Políticas es **PRESCRIPTIVA.**
* **LA CONSTRUCCIÓN DE POLITICAS generalmente IMPLICA RESOLVER problemas complicados de optimización.**

## Evaluación:

La evaluación puede ser difícil si los datos no proporcionan realmente la verdad fundamental, por que se requiere datos de gran calidad y recientes para usarlo en nuestros modelos predictivos, para ver si habíamos predicho eventos que sucedieron.

## Deployment (Implementación):

* Esto es como “¡¡Hey!! convertí plomo en oro”.
* El 95% de los proyectos se detienen en implementación, ya que nadie planea desplegarlo, a menos que sea realmente impresionante.
* Se debe tener una línea de tiempo realista, y agregar varios meses a esa línea de tiempo, debido a que los objetivos pueden cambiar, los datos pueden cambiar, como la reacción a su nueva política, ya que se puede esperar algo que no se esperaba.
* Será necesario actualizarlo y mejorarlo. Cuando se hace esto, no es una cosa de un tiro.