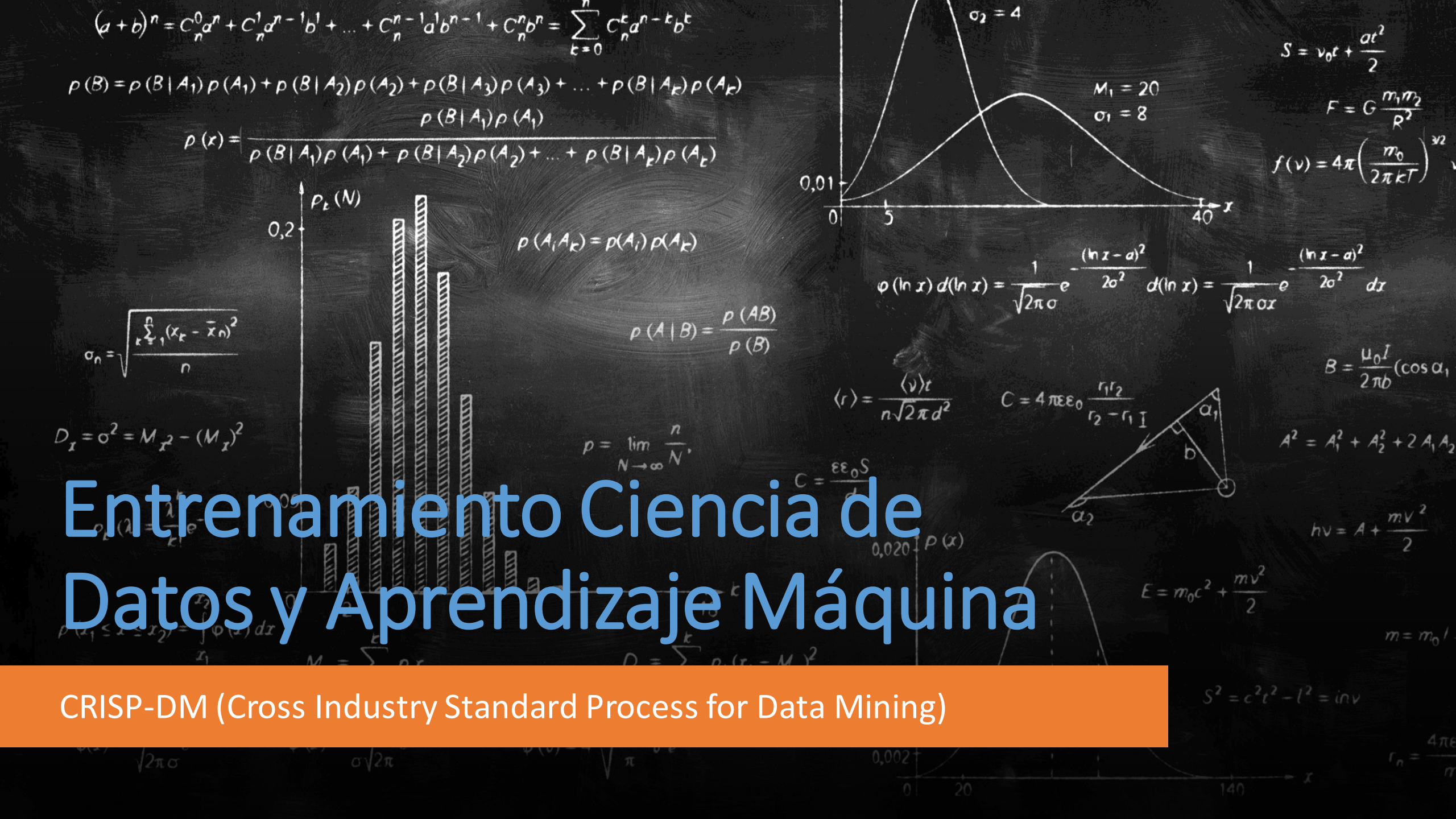
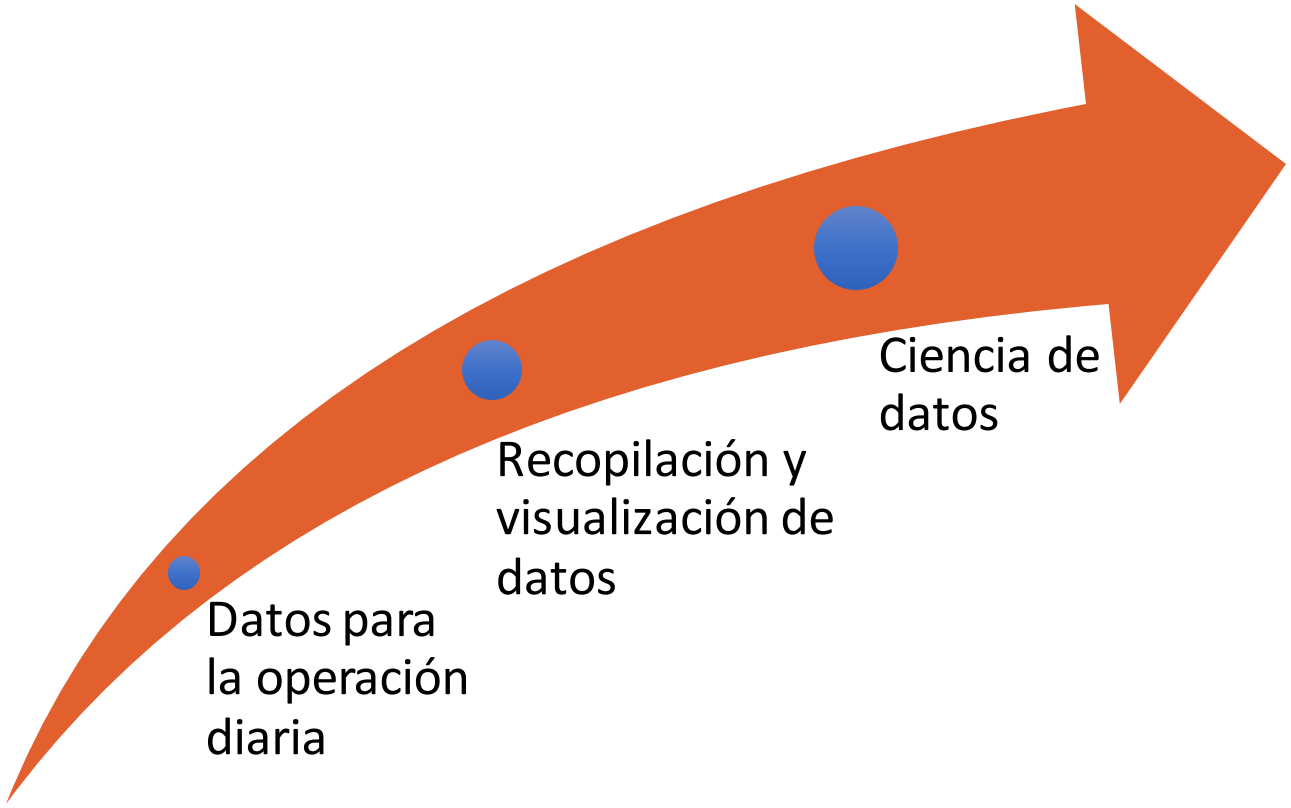


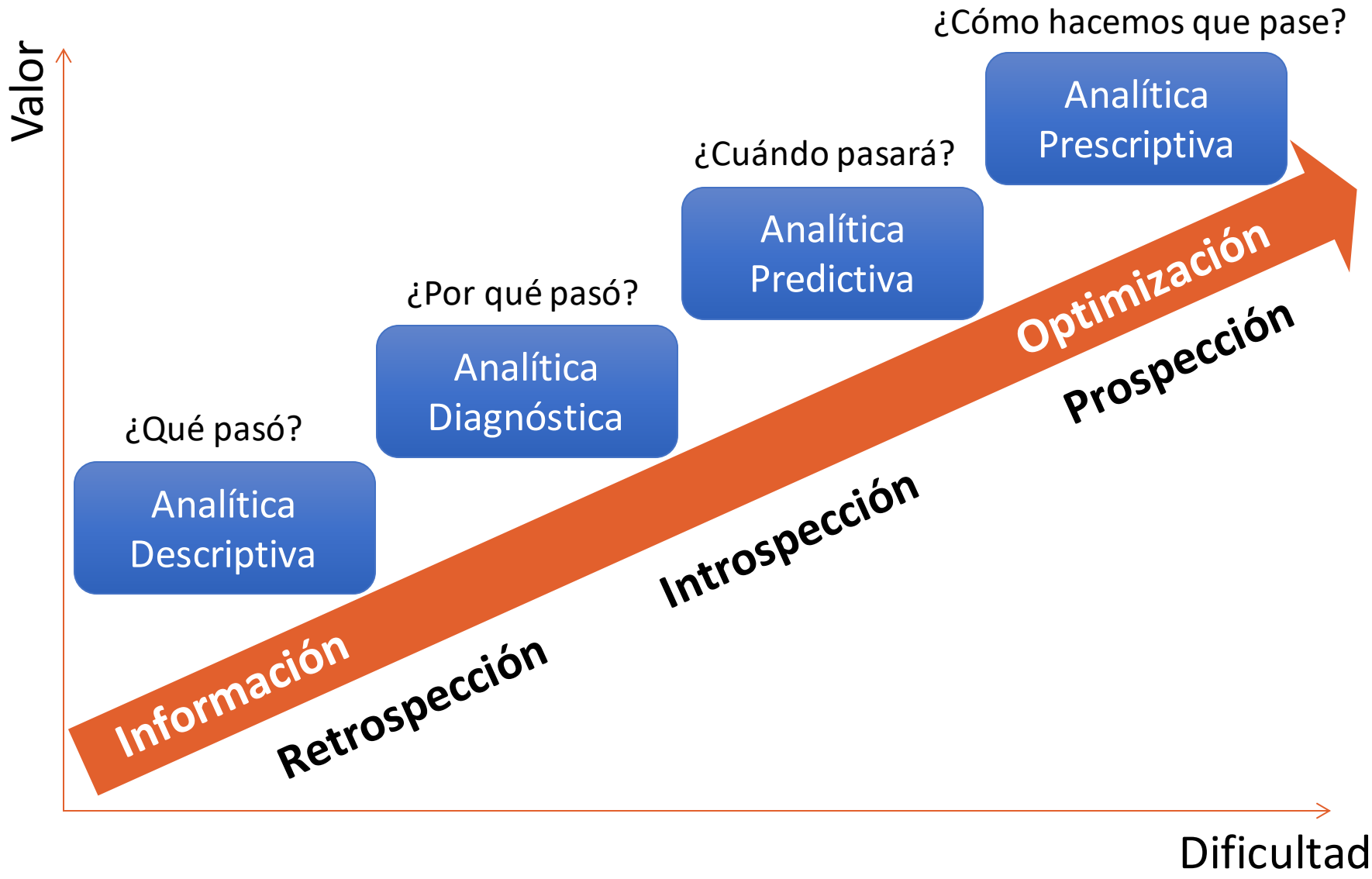
Entrenamiento Ciencia de Datos y Aprendizaje Máquina

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

Evolución de los datos



Evolución de las preguntas



Big Data

La Ciencia de Datos NO es Big Data.

El Big Data se describe en términos de:

- Volumen: enormes cantidades de datos estructurados, no estructurados y semiestructurados
- Variedad: gran diversidad en el tipo de datos, tales como correos, tweets, audio, videos, etc.
- Velocidad: respuestas rápidas para obtener la información necesaria en el tiempo preciso

Big Data

El término Big data se refiere a información que no puede ser procesada o analizada utilizando métodos tradicionales. Una base relacional sería muy lenta y costosa.

Big data requiere de la ciencia de datos para transformar la información en conocimiento

CRISP-DM

CRISP-DM es el acrónimo del inglés para Cross-Industry Standard Process for Data Mining.

Como su nombre lo indica es una metodología que proporciona una forma estructurada para resolver un problema analítico dentro de la industria.

Esta metodología divide el ciclo completo de minería de datos en 6 etapas. Con lo que ayuda a identificar puntos clave de un proyecto.

CRISP-DM

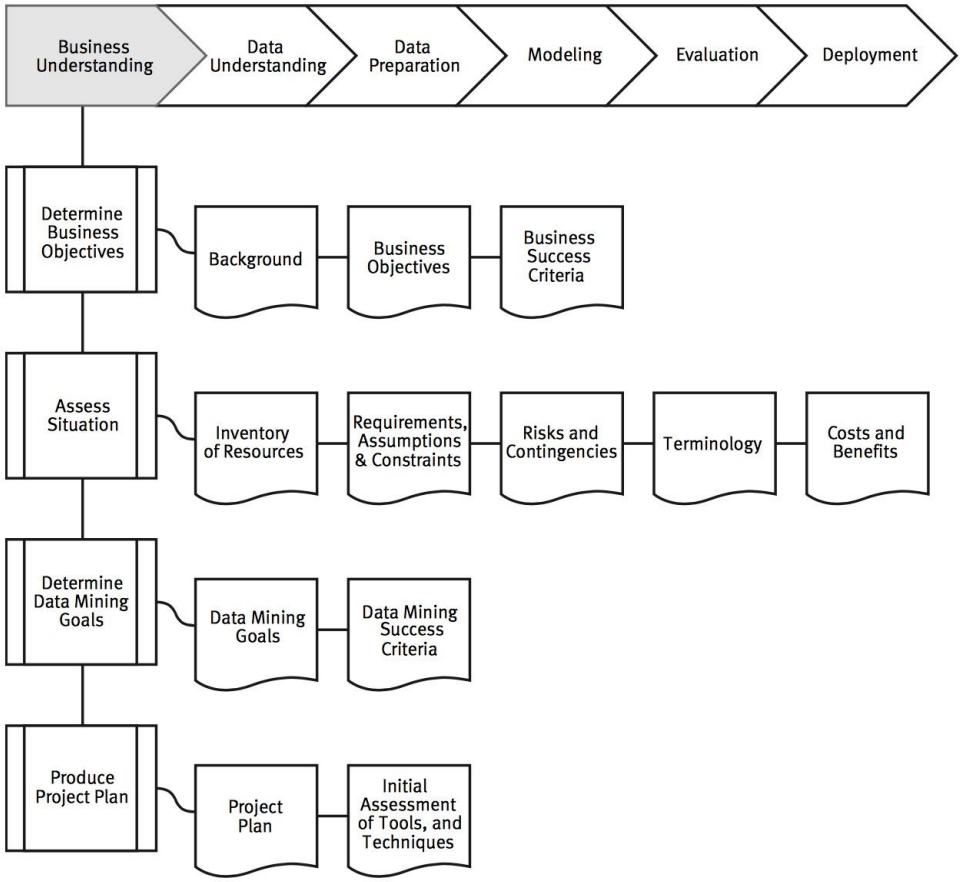
Business Understanding

En esta etapa se deben determinan los objetivos del negocio:

- Se plantea la problemática del negocio
- Se traduce del lenguaje de negocio a un problema de datos
- Se define el plan de trabajo (estimación de tiempos y recursos)

Comprensión del negocio

- Los problemas nunca vienen presentados como un problema de ciencia de datos.
- Replantear el problema, como problema de ciencia de datos es una de las cosas más importantes del CRISP-DM.



CRISP-DM

Data Understanding

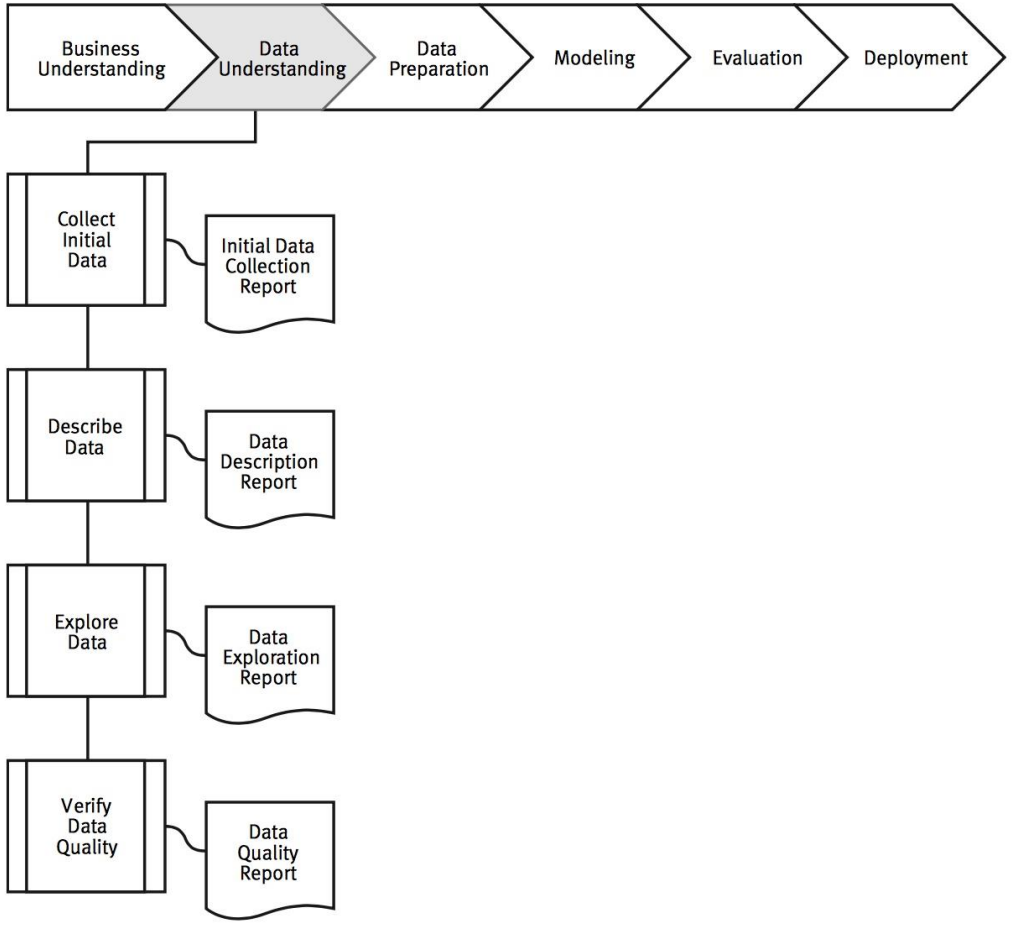
En esta etapa se junta la información necesaria para el proyecto . Se analiza el volumen, variabilidad y calidad de la misma.

Se desarrolla el análisis exploratorio de datos (EDA).

Los resultados se revisan con el negocio para adecuar los objetivos previos.

Comprensión de los datos

- Entender la limitaciones (y fortalezas) de los datos es vital.
- Rara vez los datos fueron tomados pensando en el problema.



CRISP-DM

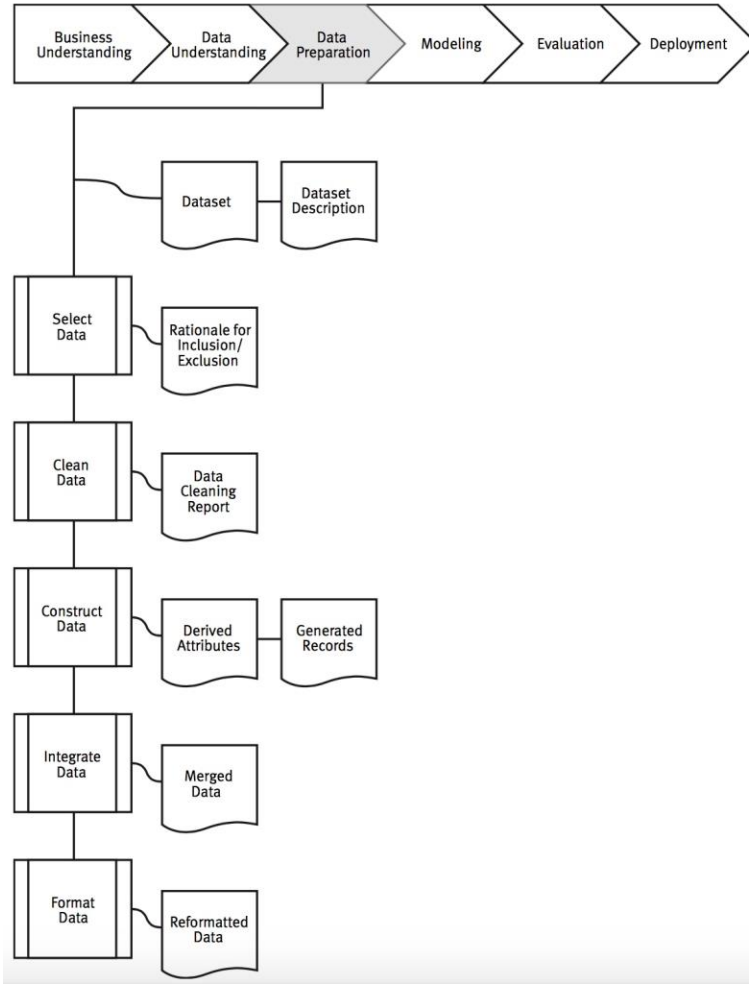
Data Preparation

En esta etapa se preparan los datos para el modelado. Las principales actividades son:

- Selección de variables
- Limpieza de variables
- Construcción de nuevas variables (feature engineering)
- Se integran todas las fuentes
- Se da formato a los datos

Preparación de los datos

- 80% es limpieza de datos…
- 20% es quejarse de la limpieza de datos



CRISP-DM

Modeling

En esta etapa se definen y entrenan diversos modelos, tanto estadísticos como de aprendizaje máquina.

Se verifican las métricas de cada modelo y, de ser necesario, se itera con la etapa anterior para mejorar los modelos.

Se elige el modelo final a evaluar.

Modelado

- Se seleccionan y aplican varias técnicas de modelado y sus parámetros se calibran a valores óptimos.

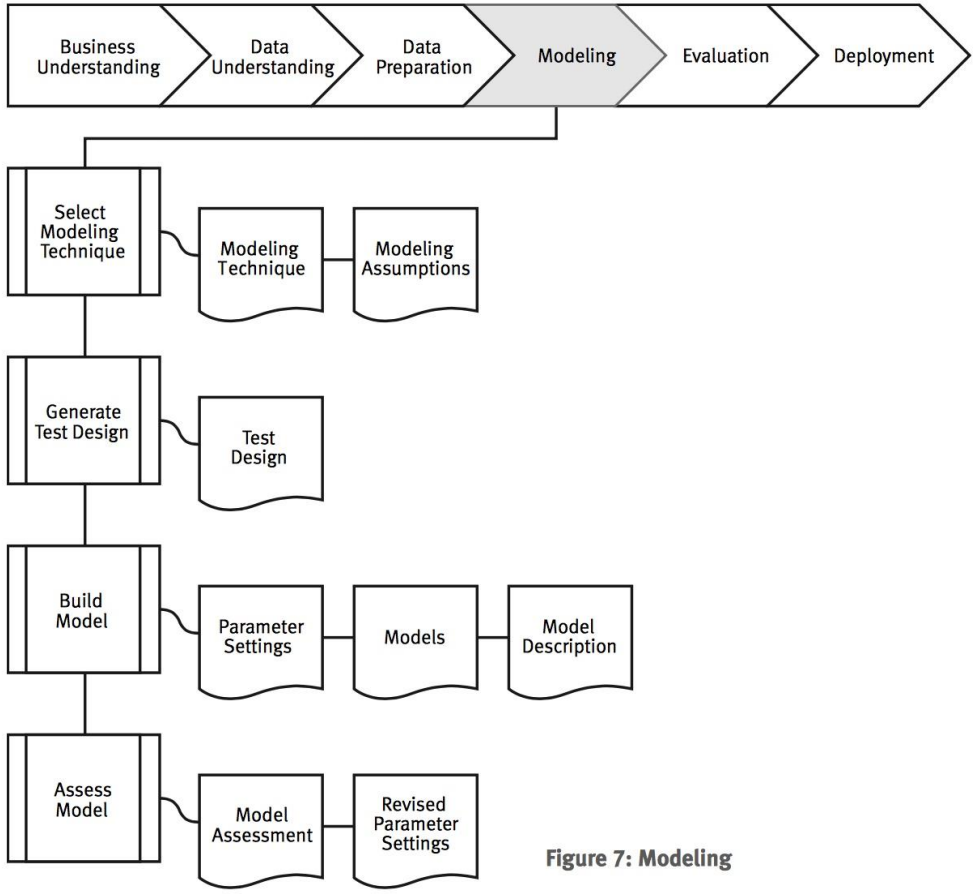


Figure 7: Modeling

CRISP-DM

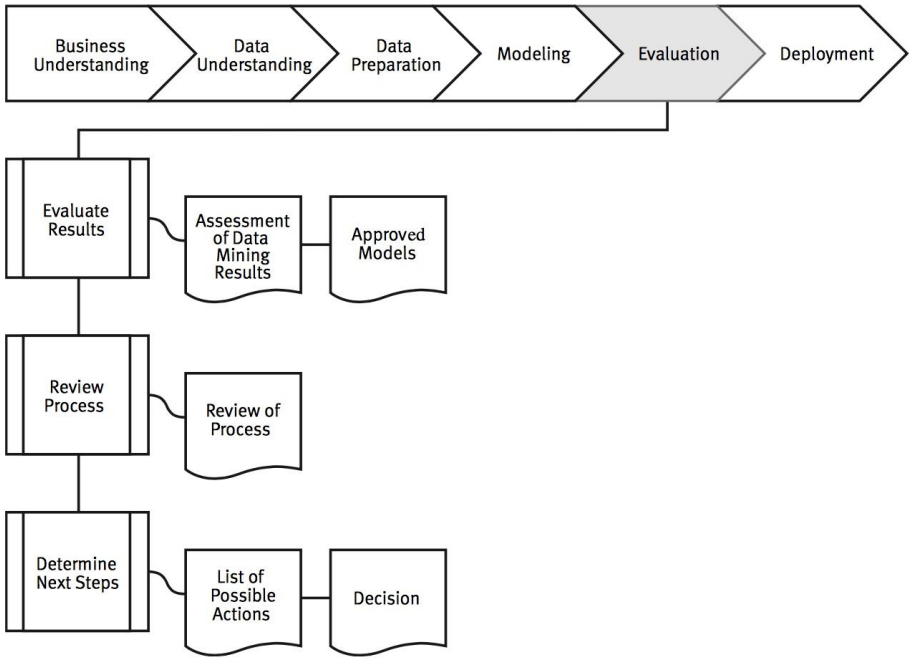
Evaluation

En esta etapa se evalúa el modelo seleccionado en la etapa anterior con datos no utilizados previamente.

Se reportan los resultados con el negocio para validar que la información obtenida hace sentido al negocio. En caso contrario se itera nuevamente el ciclo.

Evaluación

- Es importante evaluar el modelo (o modelos) a fondo y revisar los pasos ejecutados para crearlo(s), para estar seguro de que logra(n) adecuadamente los objetivos de negocio



CRISP-DM

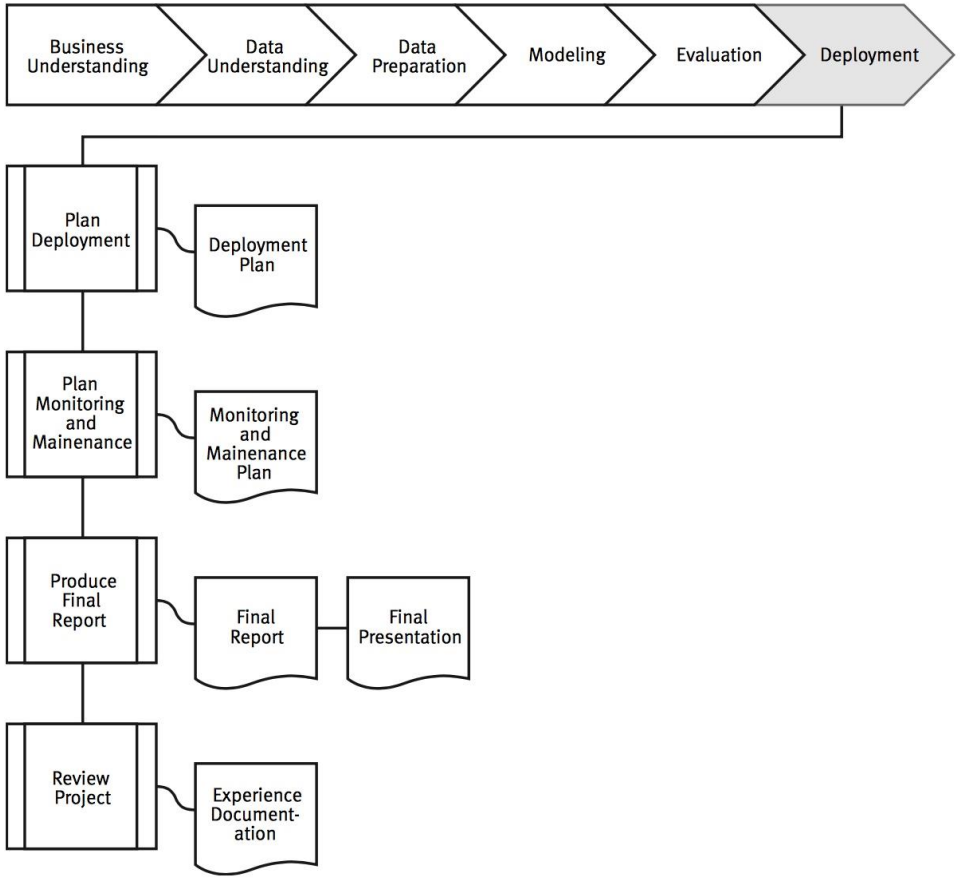
Deployment

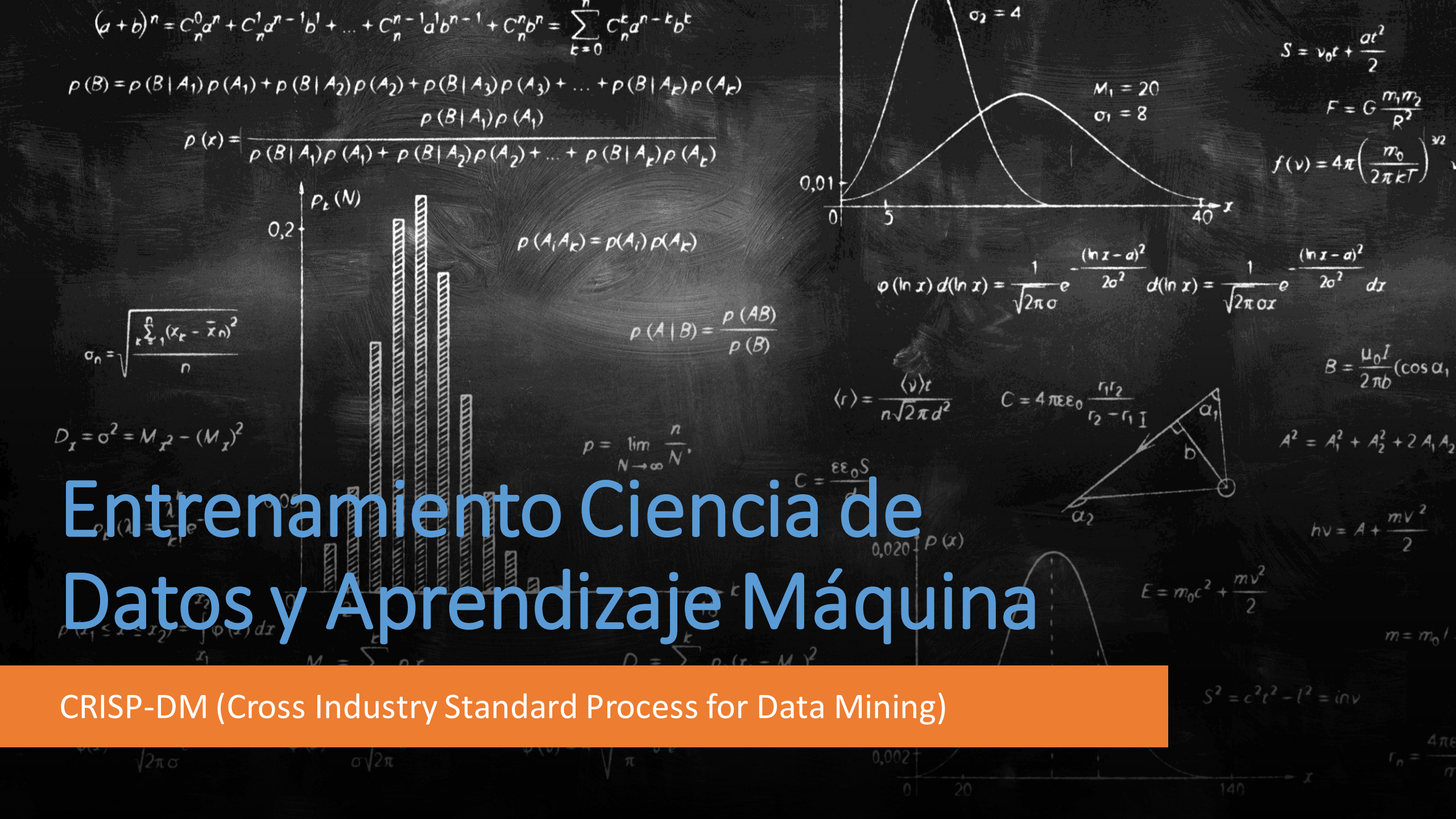
Una vez se obtiene el visto bueno por parte del negocio, se implementa el modelo obtenido en producción para su uso dentro del negocio.

Debido a cambios en las dinámicas de los negocios, es necesario evaluar de forma periódica los modelos en producción e iterar este proceso.

Implantación

- ¿Cómo impactar?
- ¿Cómo implementar?
- ¿Cómo mantener funcionando?
- El modelo no es lo que el científico de datos diseña, es lo que el ingeniero construye.





Entrenamiento Ciencia de Datos y Aprendizaje Máquina

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