09481: Inteligencia Artificial

Breyner Posso, Ing. M.Sc.
breyner.posso1@u.icesi.edu.co

Programa de Ingeniería de Sistemas.
Departamento TIC.
Facultad de Ingeniería.
Universidad Icesi.
Cali, Colombia.

Agenda

- Metodología CRISP-DM.

Metodología CRISP-DM

CRoss-Industry Standard Process for Data Mining

CRISP-DM 1.0: https://the-modeling-agency.com/crisp-dm.pdf

¿Qué es una metodología?

- La **metodología** hace referencia al conjunto de procedimientos racionales utilizados para alcanzar el objetivo o la gama de objetivos que rige una investigación científica o tareas que requieran habilidades, conocimientos o cuidados específicos.
- Con frecuencia puede definirse la metodología como el estudio o elección de un método pertinente o adecuadamente aplicable a determinado objeto.

CRISP-DM

- El desarrollo de proyectos de analítica es diferente al ciclo de desarrollo de proyectos de software.
- CRISP-DM:
  - Metodología: identifica 6 fases, sus tareas y dependencias.
  - Modelo del proceso: definición de ciclo de vida iterativo de la minería de datos.
- Puede ocurrir que se acabe una iteración sin haber resuelto la pregunta o problema original.
- IBM SPSS incorpora una herramienta de gestión de un proceso CRISP-DM.

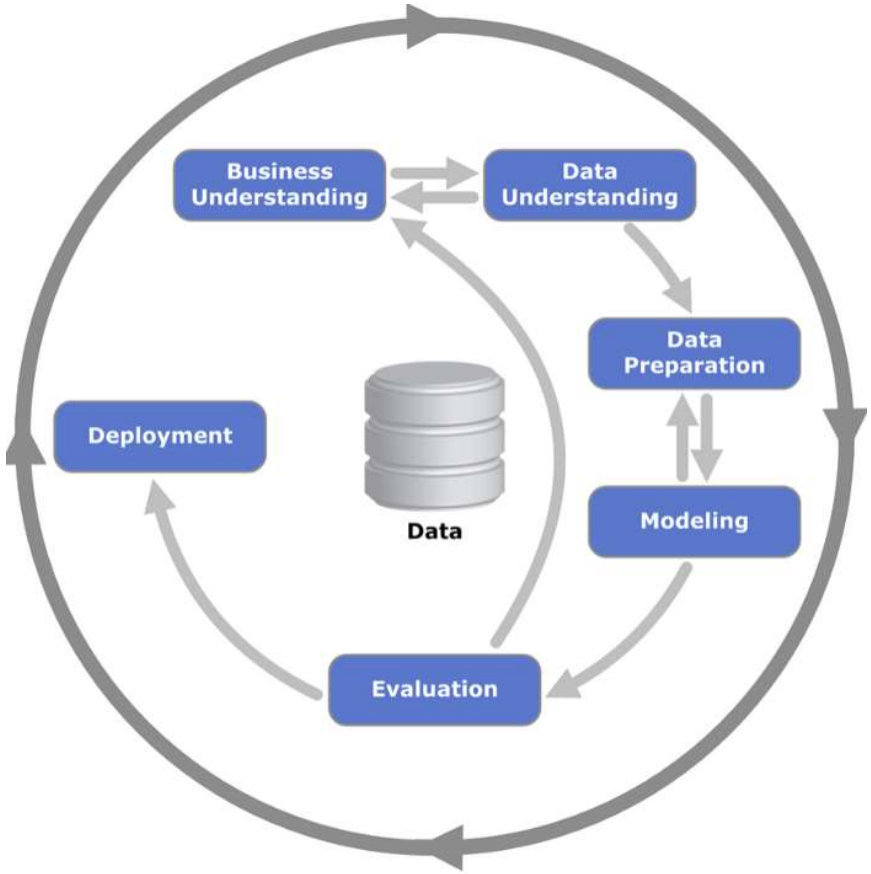


Figura: fases del modelo de referencia CRISP-DM. Fuente CRISP-DM 1.0.

https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html

¿Cuál es la historia de CRISP-DM?

- Fue un esfuerzo financiado por la Comunidad Europea para desarrollar un marco de trabajo unificado para tareas de minería de datos.
- La iniciativa comienza en 1996 liderada por Daimler Chrysler (en ese entonces Daimler-Benz), SPSS (en ese entonces ISL) y NCR.
- Se desarrolla y refina a través de una serie de talleres entre 1997 y 1999.
- Cerca de 300 organizaciones contribuyeron en el modelo del proceso.
- La versión 1.0 se publica en 1999.

¿Cuáles eran los objetivos?:

- Fomentar herramientas interoperables en todo el proceso de minería de datos.
- Eliminar la experiencia "misteriosa" y "costosa" de realizar tareas simples de minería de datos.

¿Cuáles son las ventajas que ofrece CRISP-DM?

- ✓ Permite replicar proyectos de minería de datos.
- ✓ Facilita la planeación y administración de proyectos de minería de datos.
- ✓ Facilita el ingreso de nuevos interesados en el campo.
- ✓ Promueve las buenas practicas y ayuda a obtener mejores resultados.

Pero han pasado más de 20 años … ¿si se sigue usando CRISP-DM? 🤔

www.datascience-pm.com/crisp-dm-still-most-popular/

Data Science Project Management

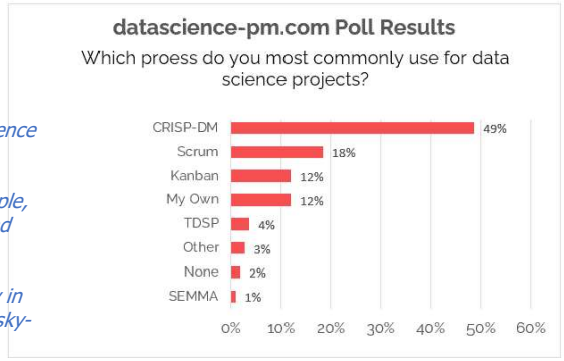
TRADITIONAL / WORKFLOW

CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects

POSTED NOVEMBER 30, 2020 JEFF SALTZ

During the past month, we conducted at poll to see what project management framework teams used to help execute their data science projects.

Based on our survey of 109 respondents, CRISP-DM was the most commonly used data science process framework (it was used by about half the respondents). This was followed by Scrum, Kanban and "my own/my organizations". The results of our survey are shown in the chart below.



Notas:

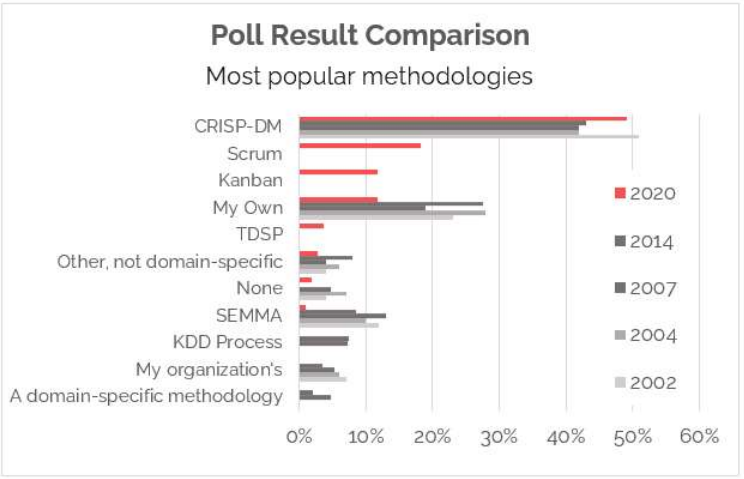
TDSP: *The Team Data Science Process (Microsoft, 2016).*

SEMMA: acrónimo de Sample, Explore, Modify, Model, and Assess (SAS, 2008).

KDD: *knowledge discovery in databases (Gregory Piatetsky-Shapiro, 1989).*

Comparing to previous surveys

The last time a survey was conducted to understand which methodology teams used for data science projects was in 2014 by KDNuggets. KDNuggets also did polls in 1010, 2007 and 2002. In comparing the results of previous surveys to our most recent survey, one of the most interesting findings from our survey is that the percentage of people using CRISP-DM has not significantly changed during the past 20 years (see the full results in the chart below).

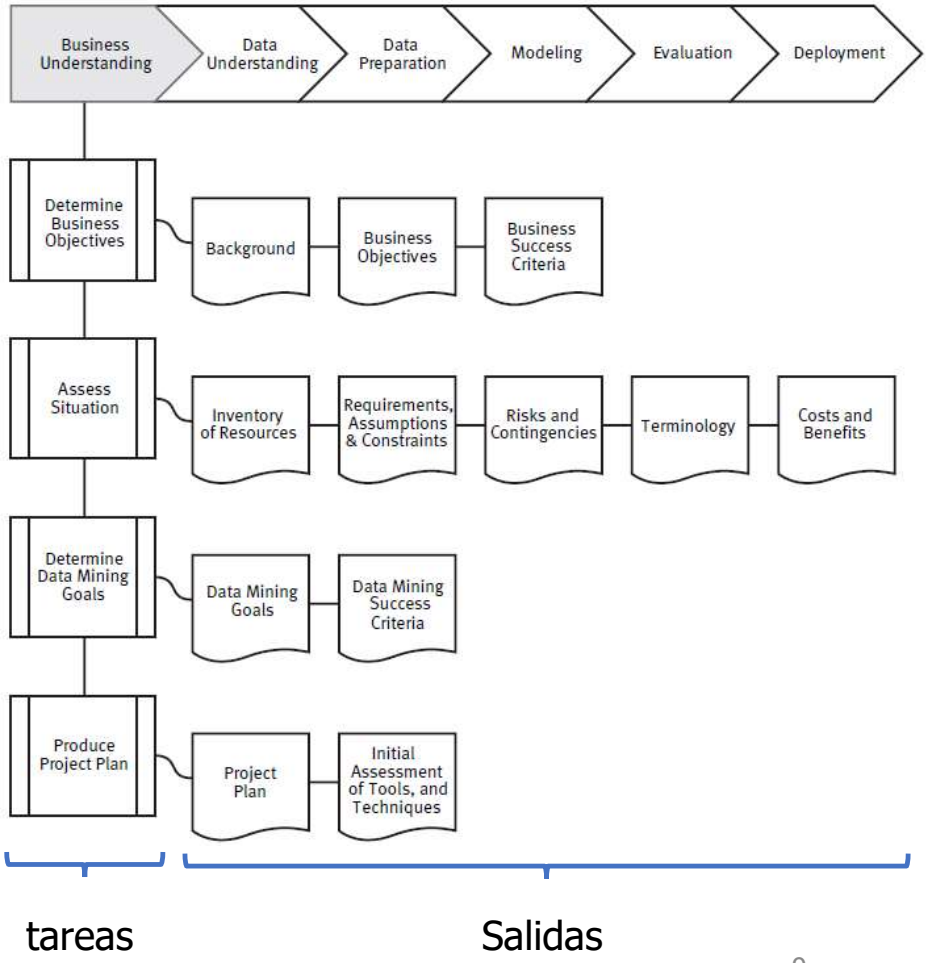


Note that our survey choices were similar to the options offered in the previous KDD polls, except we added two additional frameworks (Scrum and Kanban) and the removal of three options (Other, not domain-specific; KDD Process, My organization's)

CRISP-DM

1. Entendimiento del negocio

- ¿Cuál es el problema que se va a resolver?
- Entender los objetivos y requerimientos del proyecto desde la perspectiva del negocio.
- Luego, convertir este conocimiento en la definición de un problema de analítica y hacer un plan preliminar para cumplir los objetivos.

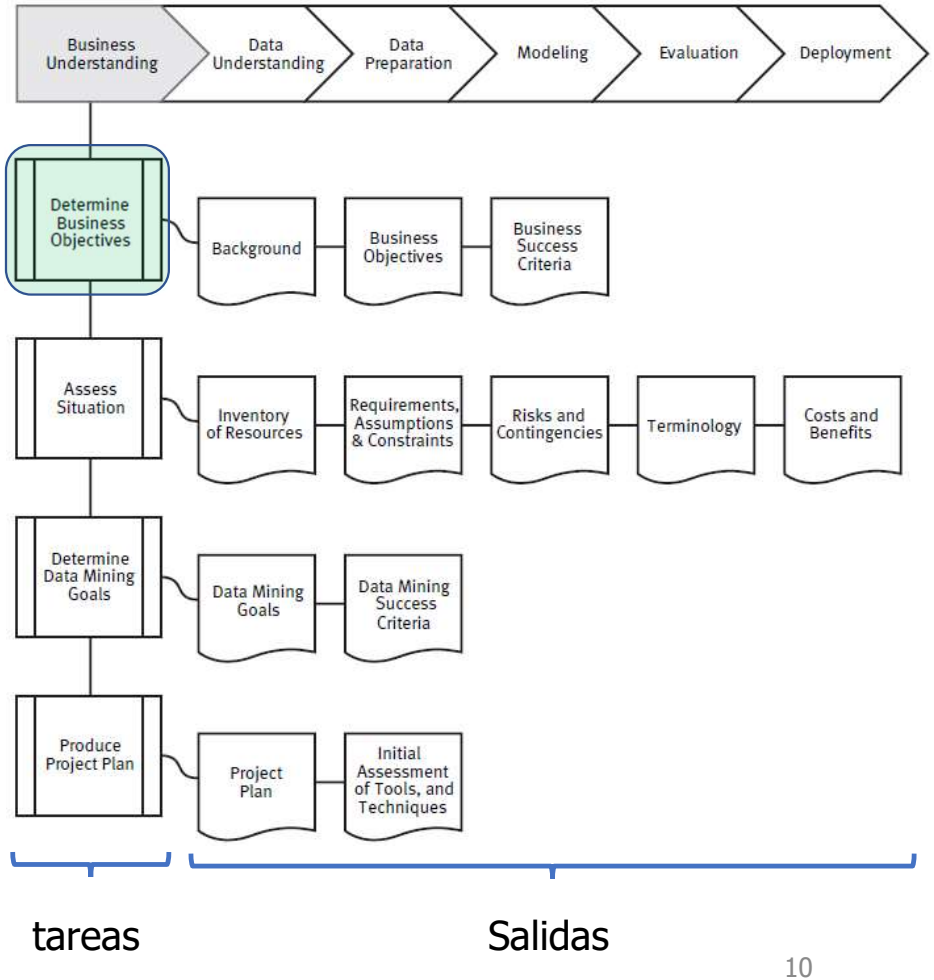


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 1****: Determinar los objetivos de negocio.*

- ¿Qué es lo que el cliente realmente desea conseguir?
- La tarea del analista es descubrir factores importantes, desde el inicio, que pueden influir en el resultado final del proyecto.
- Una posible consecuencia de saltarse este paso es invertir mucho esfuerzo para producir las respuestas correctas a las preguntas equivocadas.

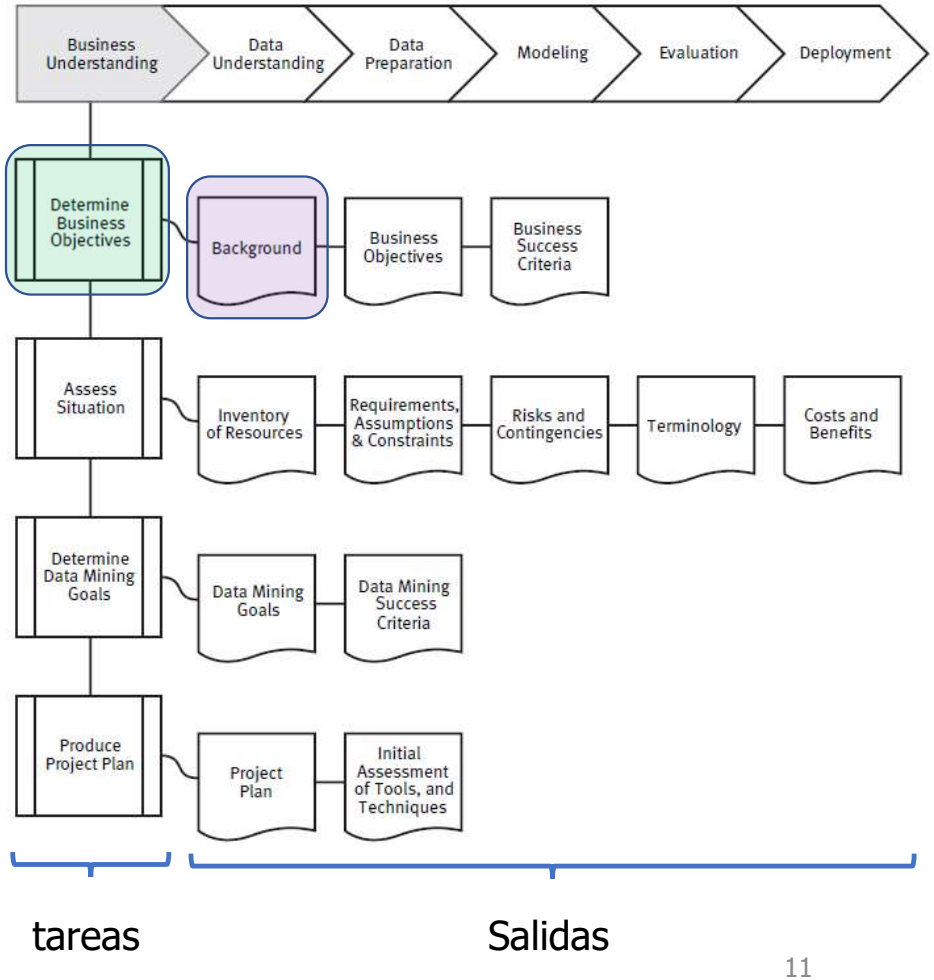


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 1****: Determinar los objetivos de negocio.*
- *Salida 1:* Background

- Recopilación de la información que se conoce sobre la situación empresarial de la organización al inicio del proyecto.

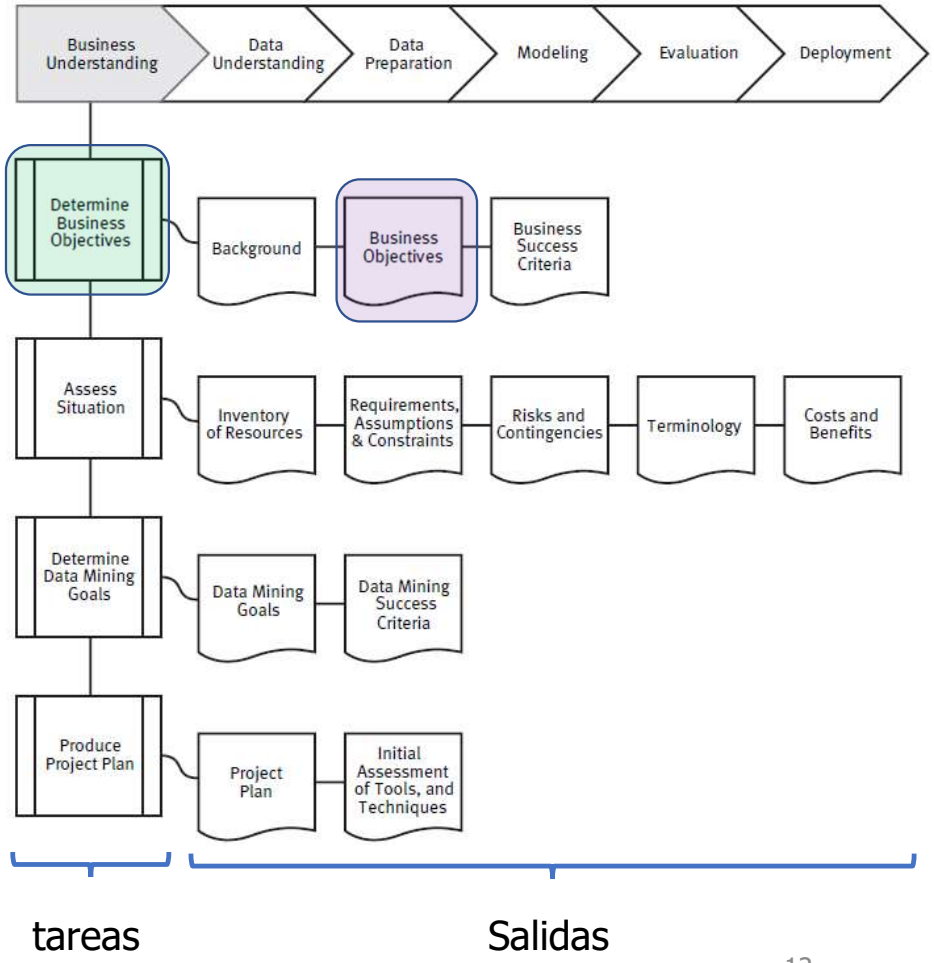


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 1****: Determinar los objetivos de negocio.*
- *Salida 2:* Business Objectives

- Descripción del objetivo principal del cliente desde una perspectiva de negocio.

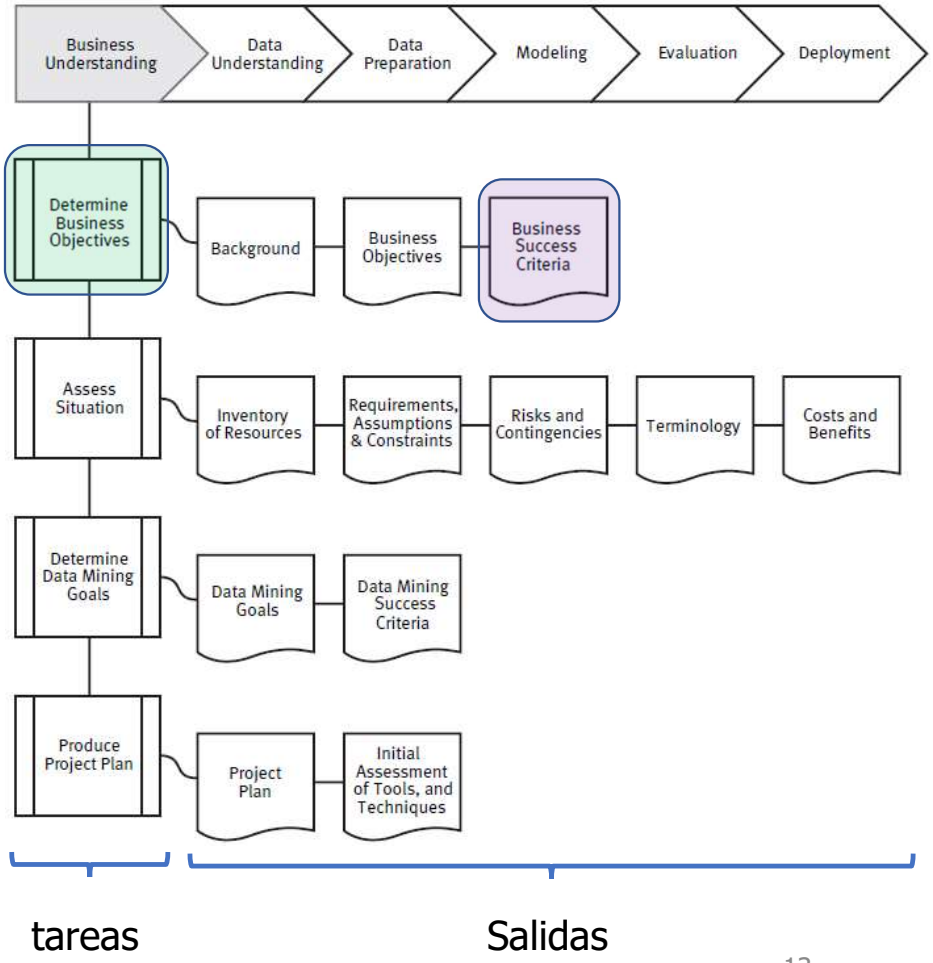


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 1****: Determinar los objetivos de negocio.*
- *Salida 3:* Business Success Criteria

- Descripción del criterio para considerar exitoso o útil el resultado del proyecto desde el punto de vista de negocio.

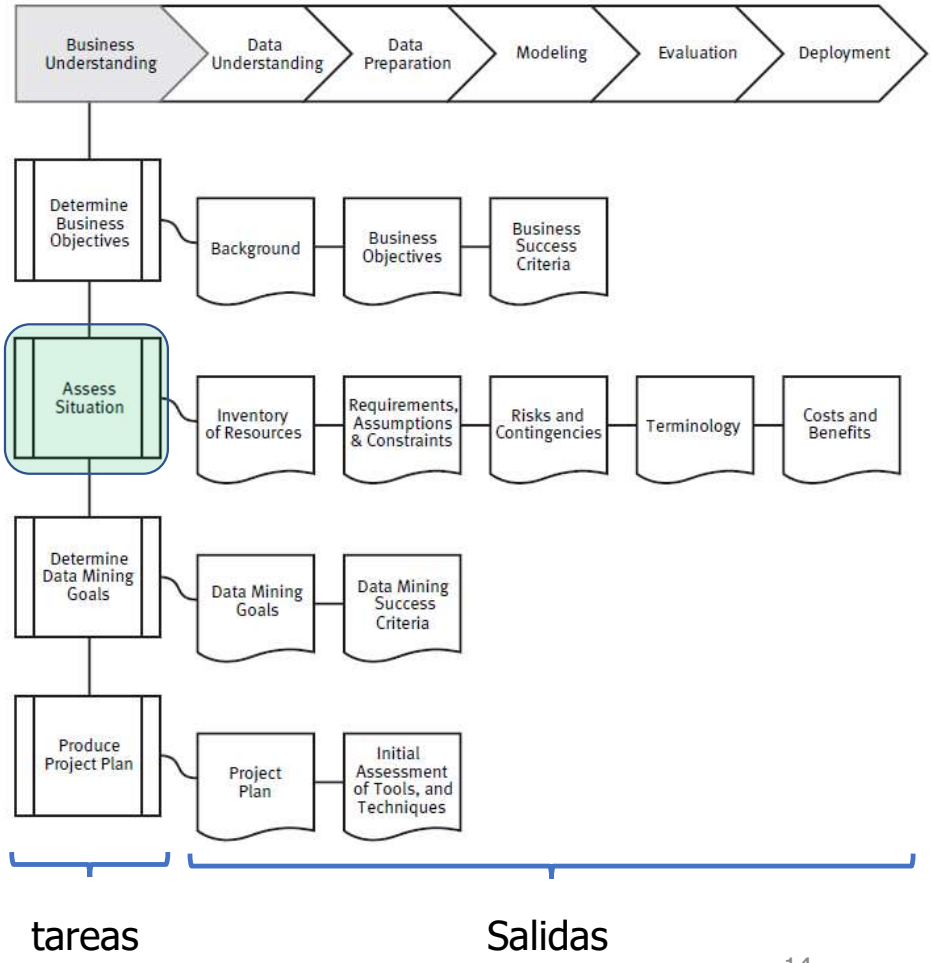


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 2**: Evaluar la situación.*

- Esta tarea implica una búsqueda de hechos más detallada sobre todos los recursos, restricciones, suposiciones y otros factores que se deben considerar al determinar el objetivo del análisis de datos y el plan del proyecto.

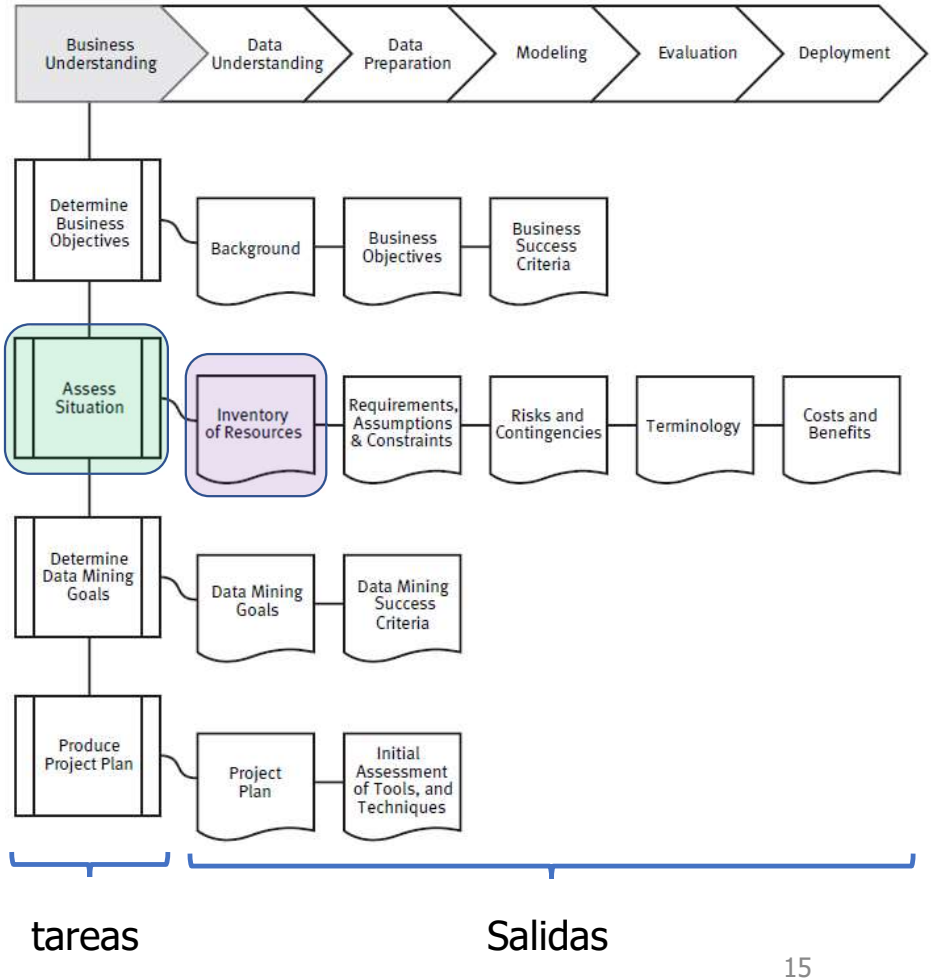


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 2****: Evaluar la situación.*
- *Salida 1:* Inventory of Resources

- Lista de todos los recursos disponibles: personal (expertos, soporte técnico, etc.), datos, recursos computacionales y software.

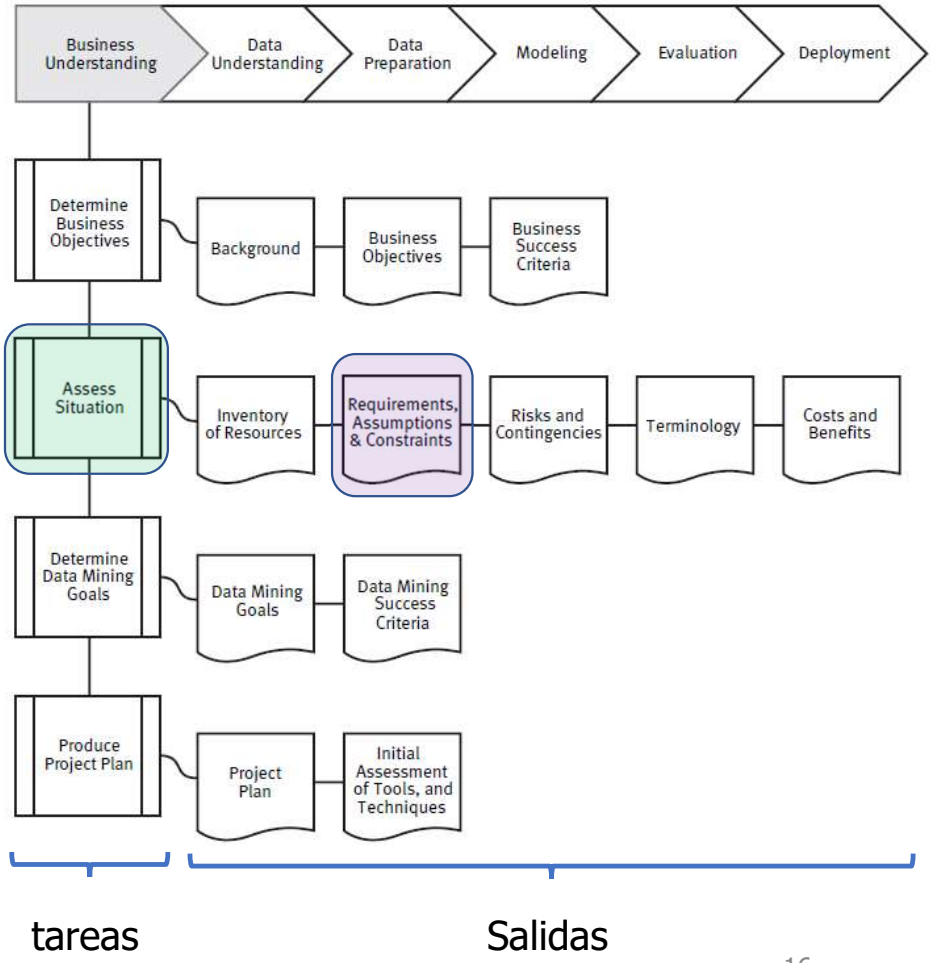


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 2****: Evaluar la situación.*
- *Salida 2:* Requirements, assumptions and Constraints

- Lista de todos los requerimientos.
- Lista de todas las suposiciones hechas para el proyecto (incluido el acceso y la utilidad de los datos).
- Lista de todas las restricciones del proyecto (disponibilidad de recursos, legales, tecnológicas, etc.).

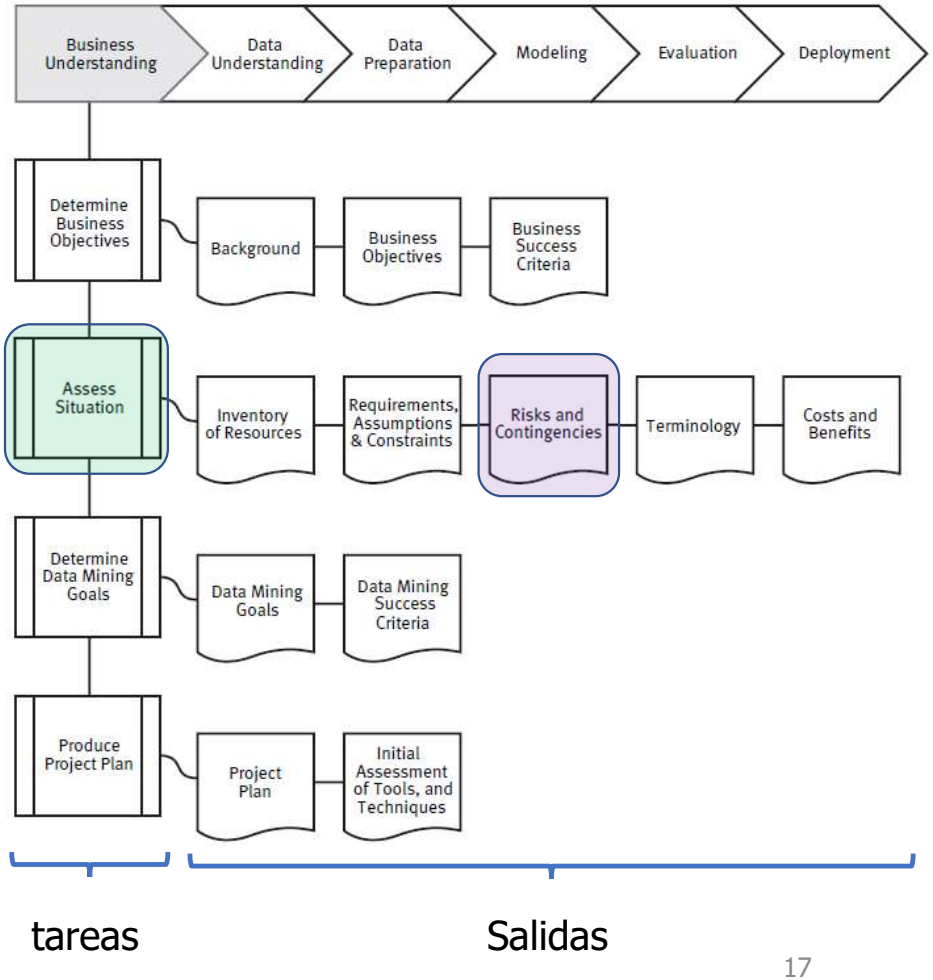


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 2****: Evaluar la situación.*
- *Salida 3:* Riesgos y contingencias

- Lista de todos los riesgos o eventos que pueden retrasar el avance del proyecto o incluso hacer que no se pueda realizar.
- Lista de planes de contingencia ante los riesgos o eventos detectados en el punto anterior.

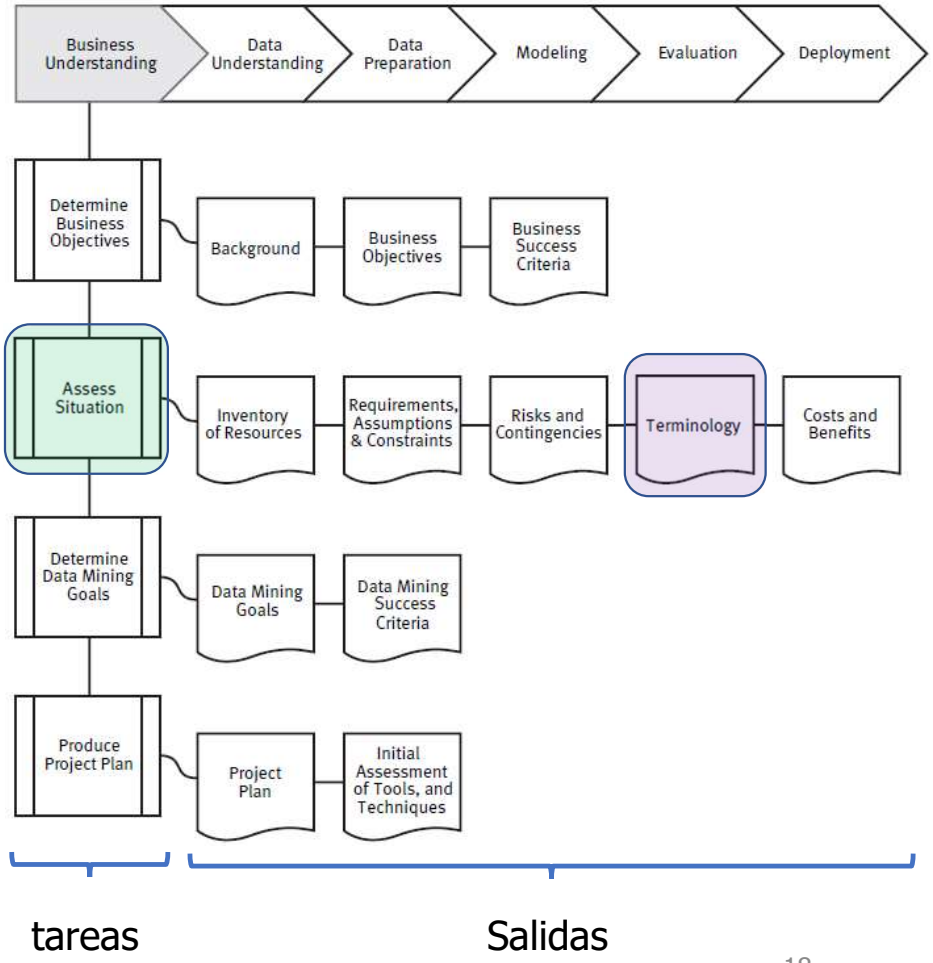


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 2****: Evaluar la situación.*
- *Salida 4:* Terminology

- Glosario de términos relevantes al proyecto: terminología relevante para el negocio y terminología importante en el proceso de analítica.

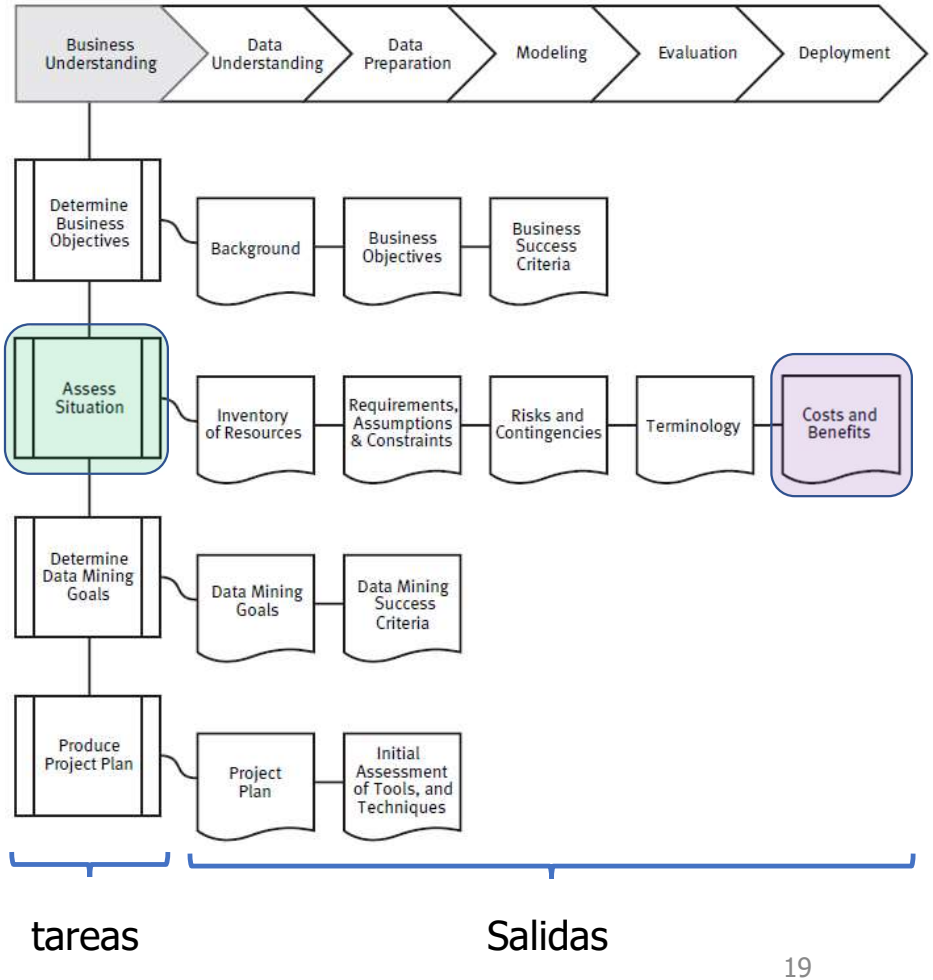


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 2****: Evaluar la situación.*
- *Salida 5:* Costs and Benefits

- Análisis de costo-beneficio: cuánto cuesta el proyecto y cuáles son los potenciales beneficios si el proyecto es exitoso.
- Este análisis debe ser lo más específico posible.

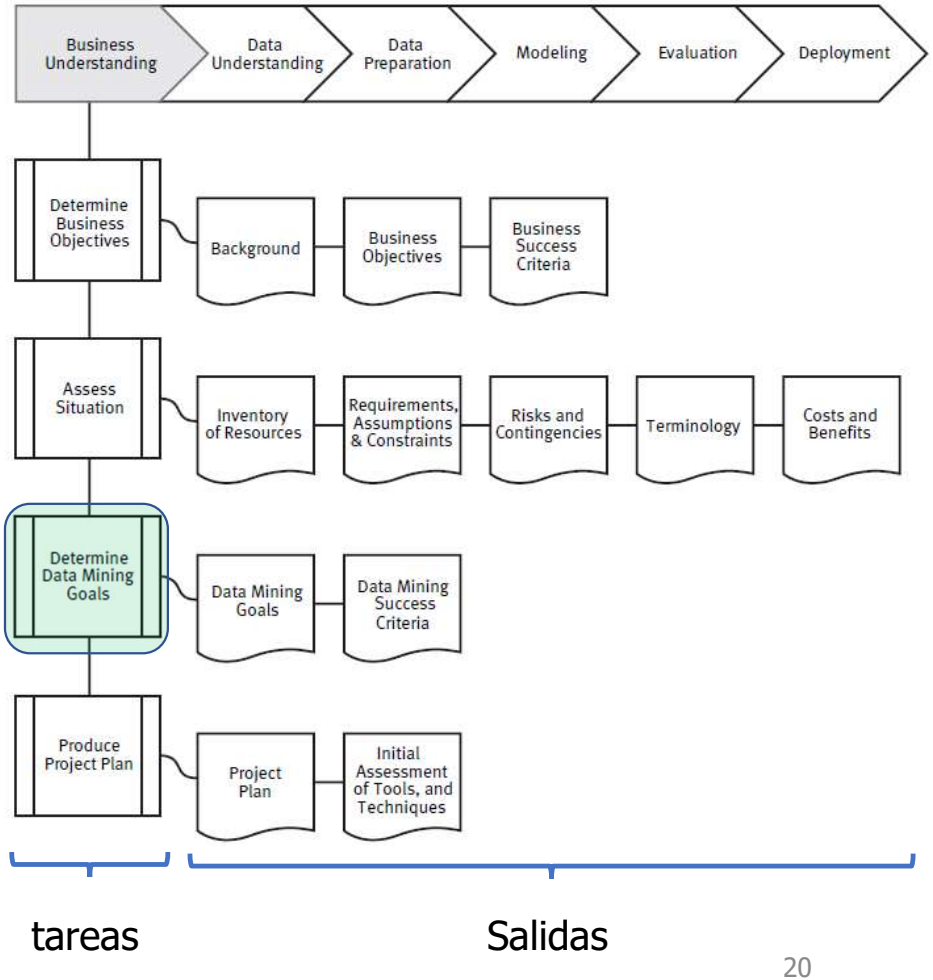


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 3****: Determinar los objetivos de la analítica.*

- Un objetivo de negocio usa terminología de negocio.
- Un objetivo de la analítica usa terminología técnica.

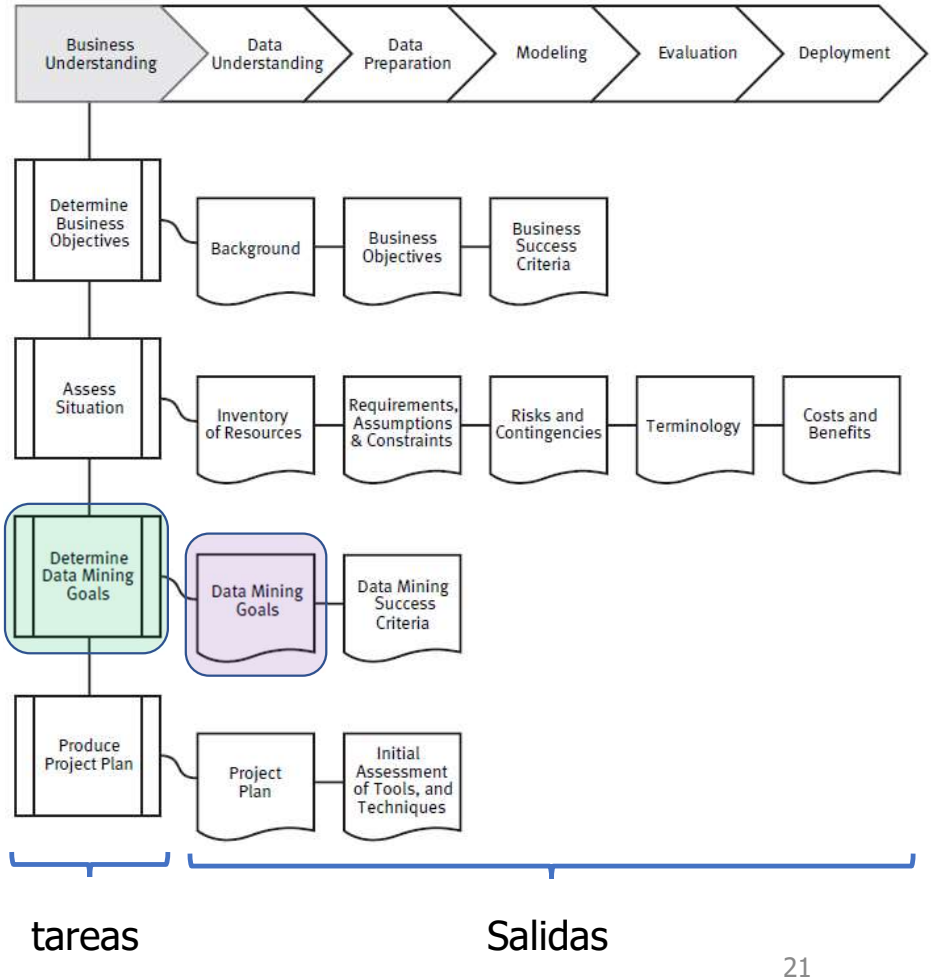


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 3****: Determinar los objetivos de la analítica.*
- *Salida 1:* Data Mining Goals

- Descripción de los objetivos de la analítica que permitirán que se cumplan los objetivos de negocio.

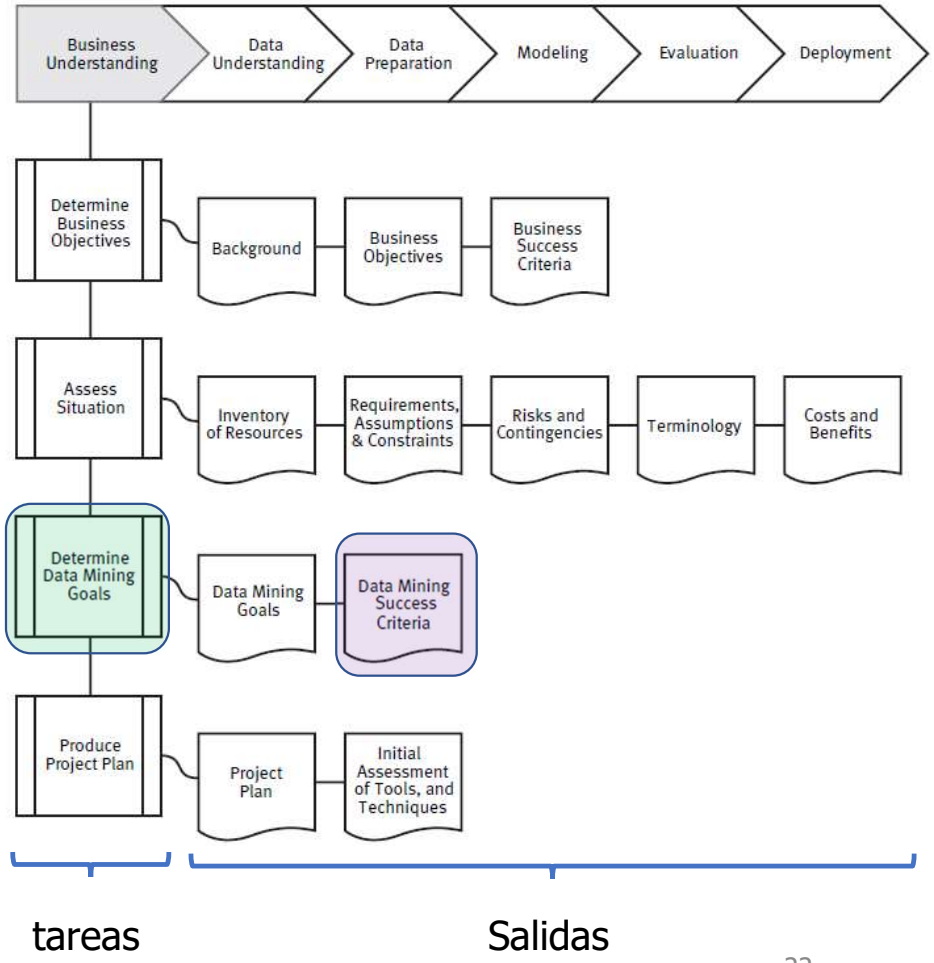


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 3****: Determinar los objetivos de la analítica.*
- *Salida 2:* Data Mining Succes Criteria

- Descripción de los criterios de éxito usando terminología técnica.
- Los criterios de éxito del proyecto pueden ser cuantitativos (cuidado!!) o cualitativos (un poco más subjetivos).

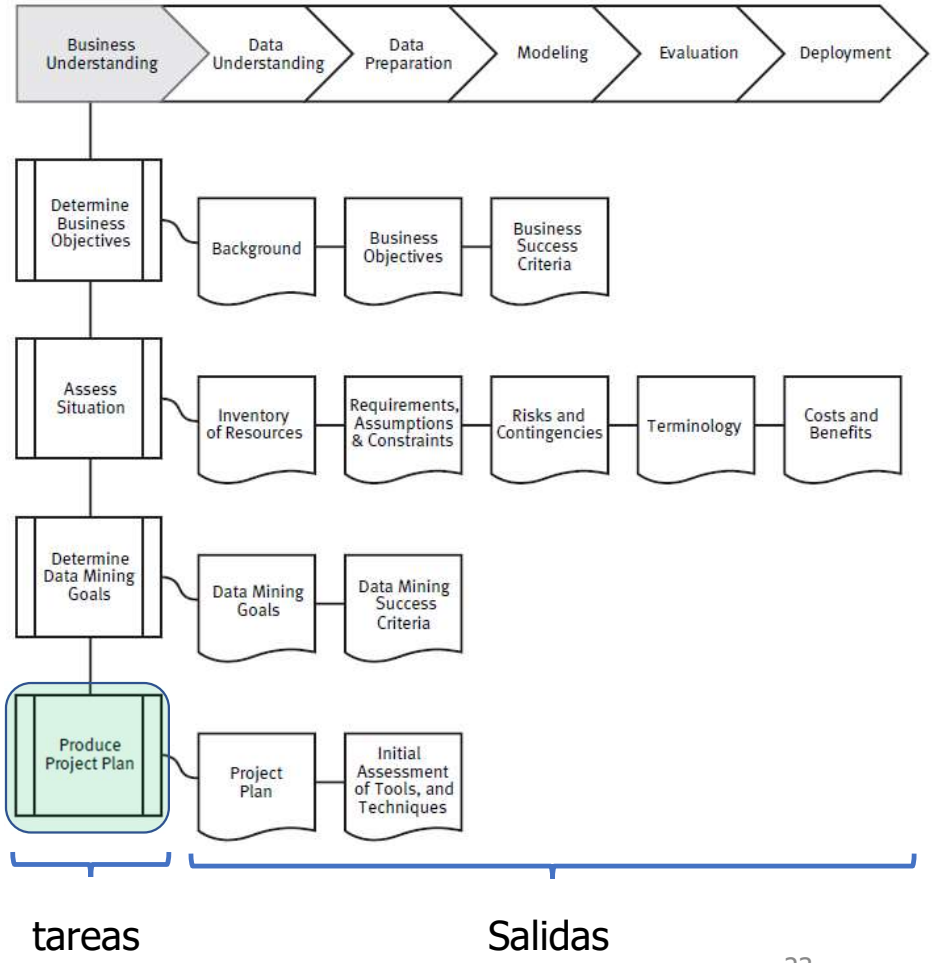


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 4**: Hacer el plan del proyecto.*

- El plan debe especificar todos los pasos que se deben llevar a cabo para alcanzar los objetivos de la analítica y por lo tanto, los objetivos de negocio.

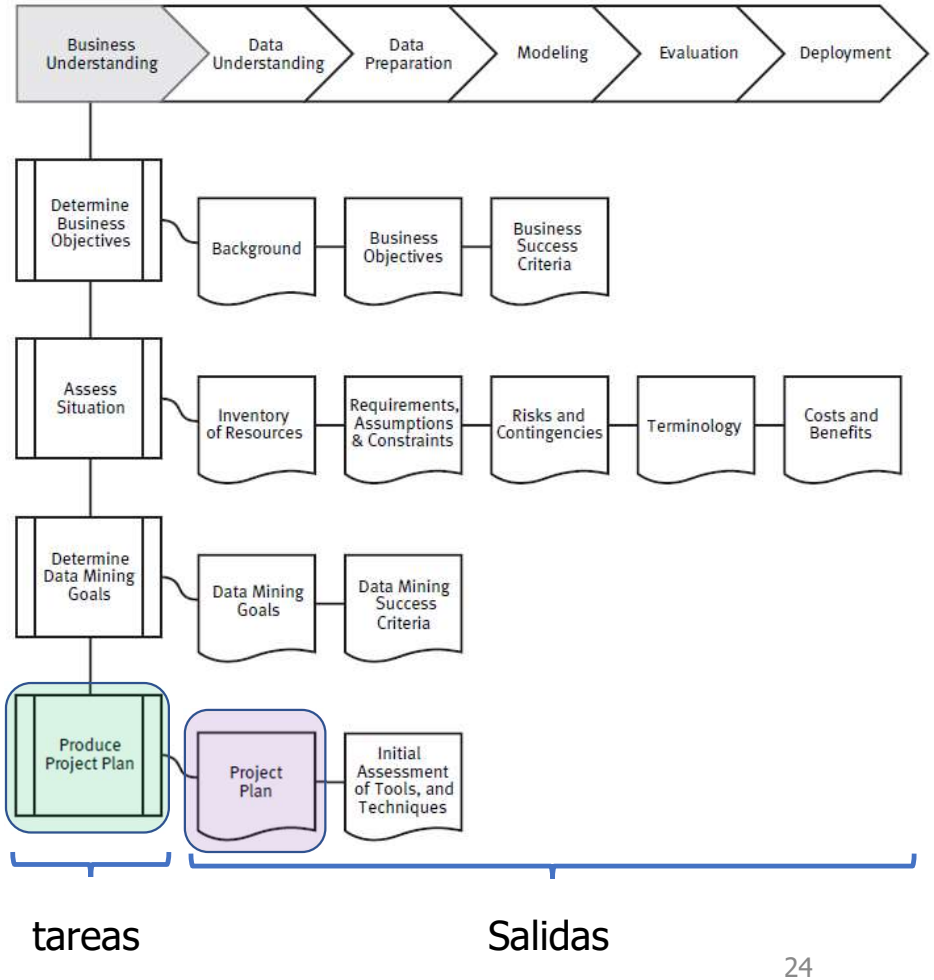


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 4****: Hacer el plan del proyecto.*
- *Salida 1:* Project Plan

- Lista de todas las etapas junto con duración, recursos, insumos, salidas y dependencias.
- Análisis de dependencias, riesgos y contingencias.
- El plan del proyecto se revisa periódicamente y se puede actualizar.

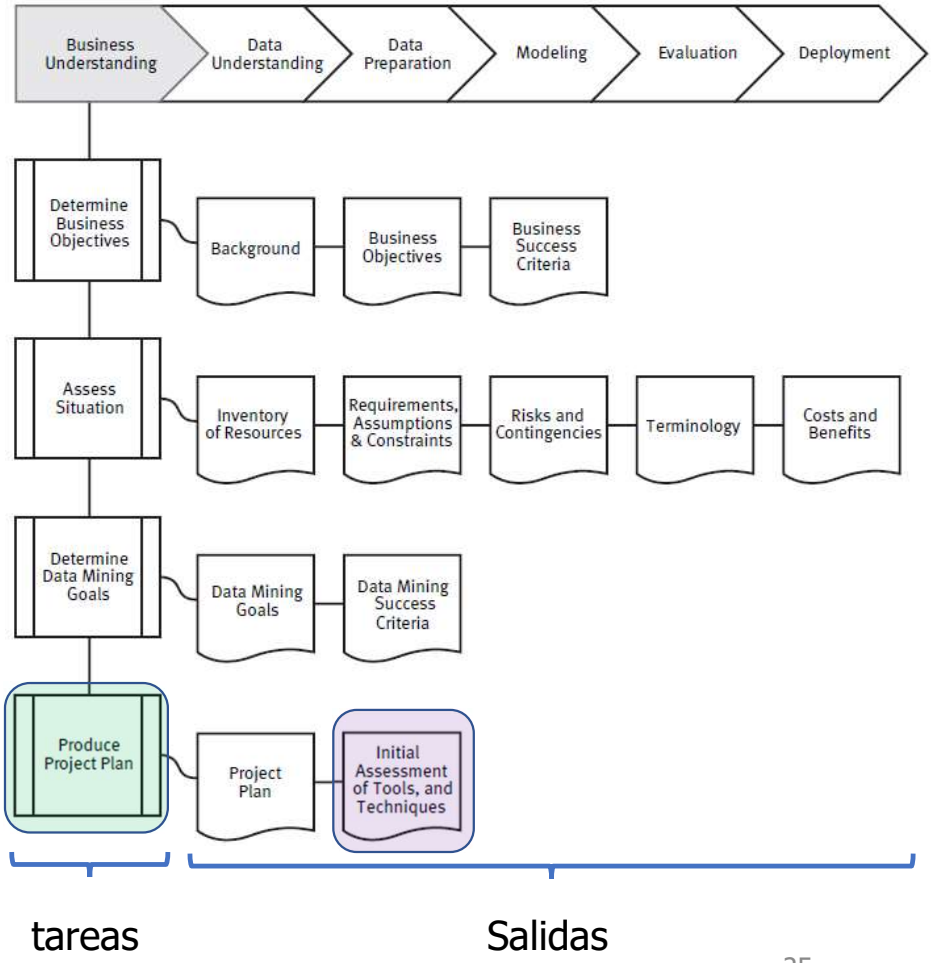


CRISP-DM

1. Entendimiento del negocio

- ***Tarea 4****: Hacer el plan del proyecto.*
- *Salida 2:* Initial Assessment of Tools and Techniques

- Evaluación inicial de herramientas y técnicas.
- Comprensión de soluciones previas, si existen.



CRISP-DM

2. Entendimiento de los datos

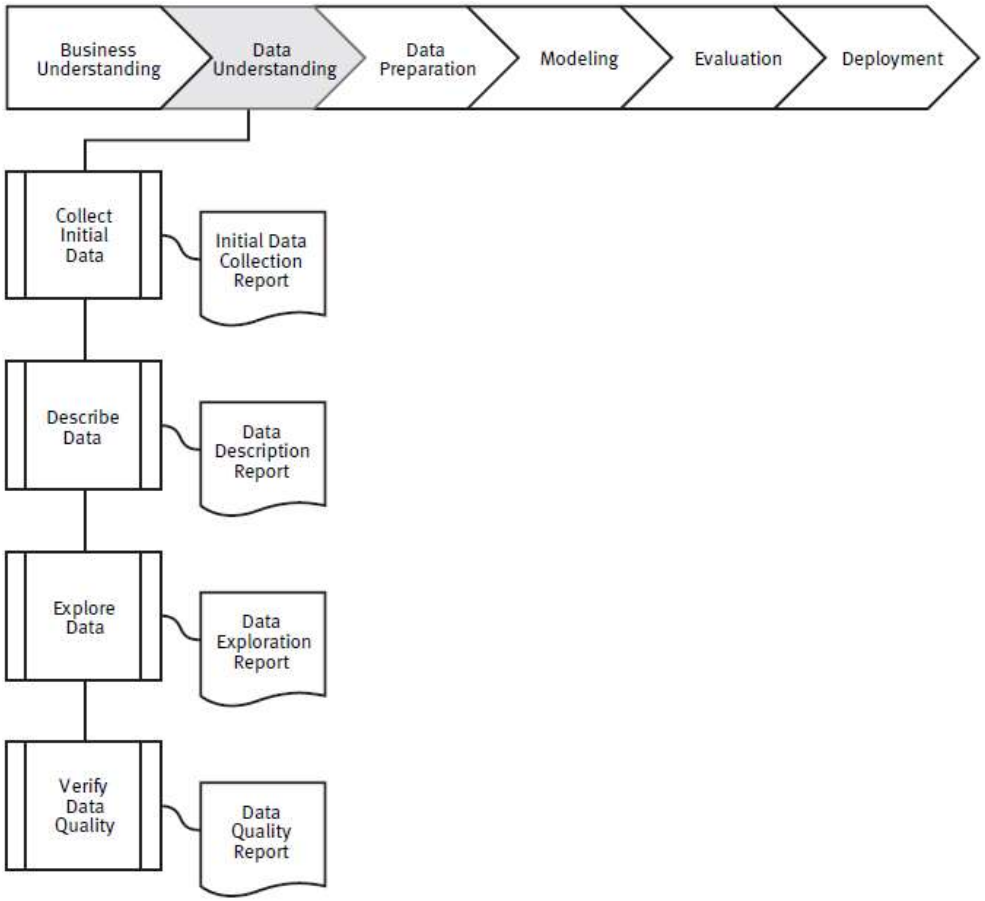
- Familiarizarse con los datos.

- ***Tarea 1****: Recolección inicial de datos.*

- ***Tarea 2****: Descripción de los datos*

- ***Tarea 3****: Exploración de datos*

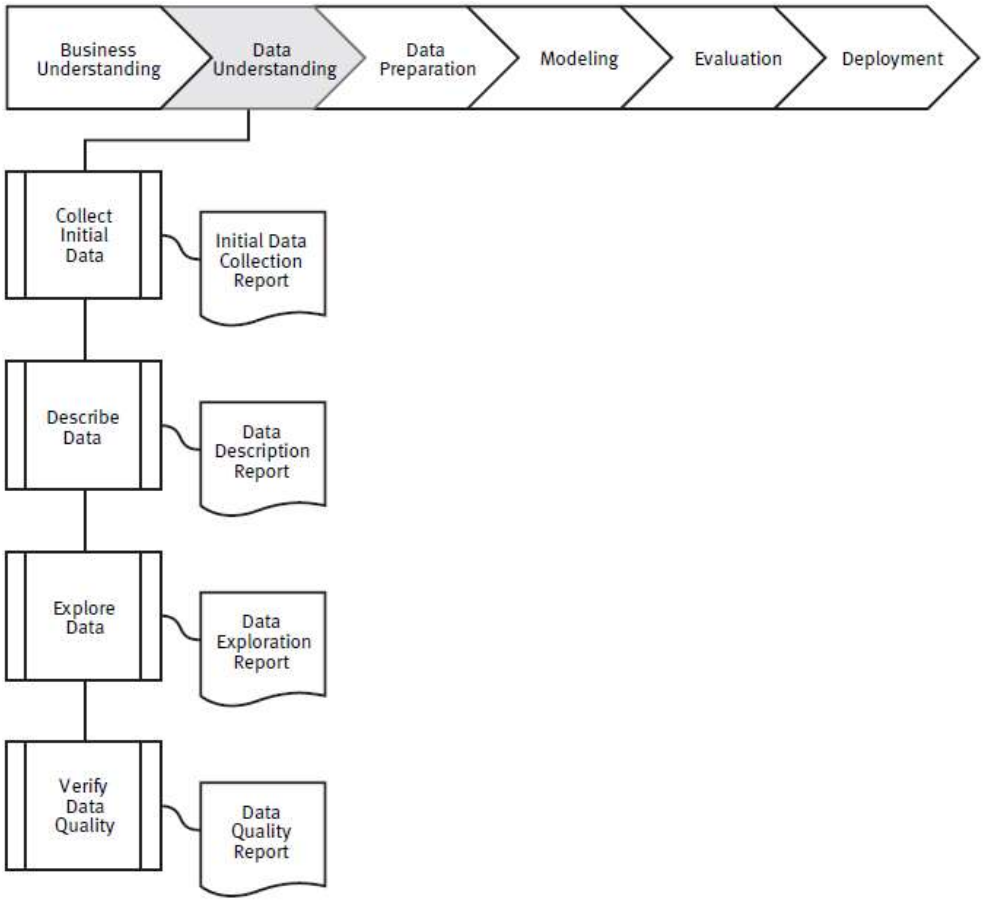
- ***Tarea 4****: Verificar la calidad de los datos*



CRISP-DM

2. Entendimiento de los datos

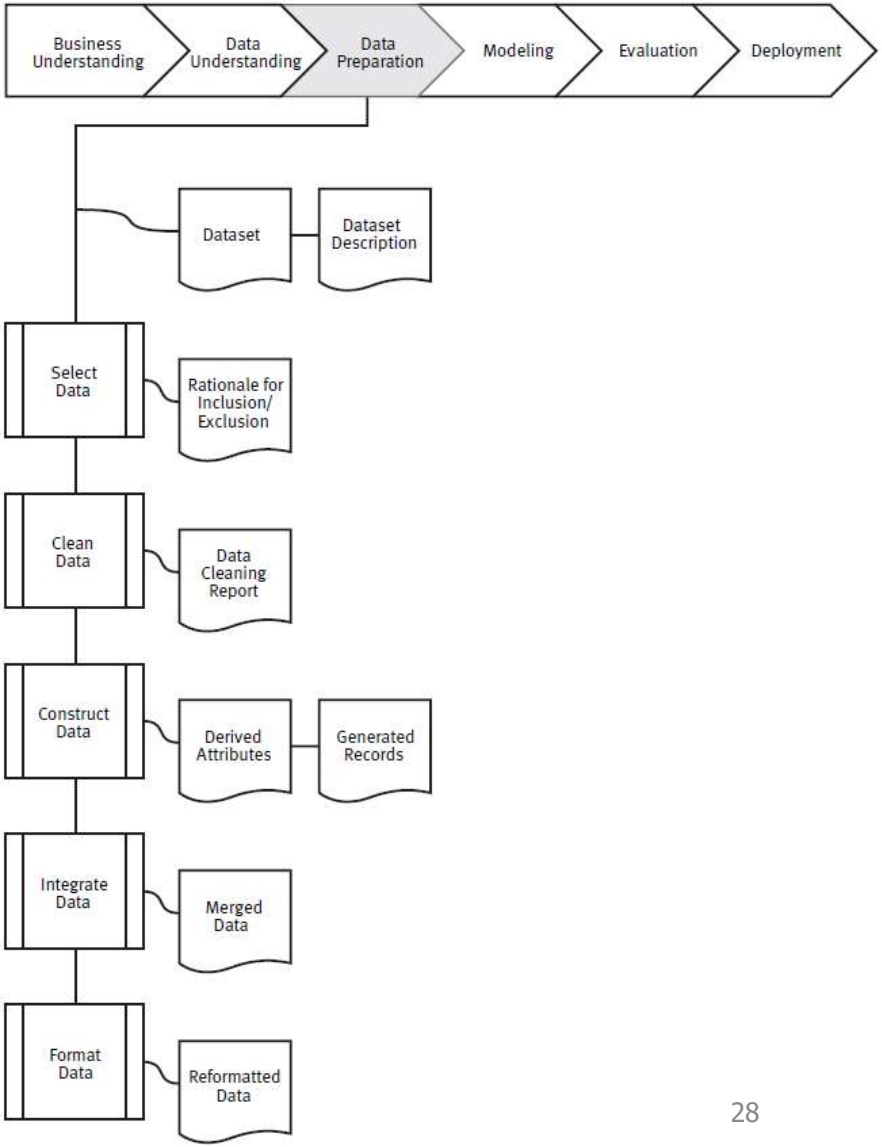
- Familiarizarse con los datos.
- Adquisición inicial: dónde, cómo, problemas y soluciones.
- Descripción: formato, cantidad, atributos, etc.
- Exploración: hacer los primeros descubrimientos. Detectar subconjuntos interesantes para formular hipótesis con respecto a la información por descubrir. Identificar fortalezas y limitaciones.
- Verificación de la calidad: identificar errores, valores faltantes o nulos. Listar posibles soluciones.



CRISP-DM

3. Preparación de los datos

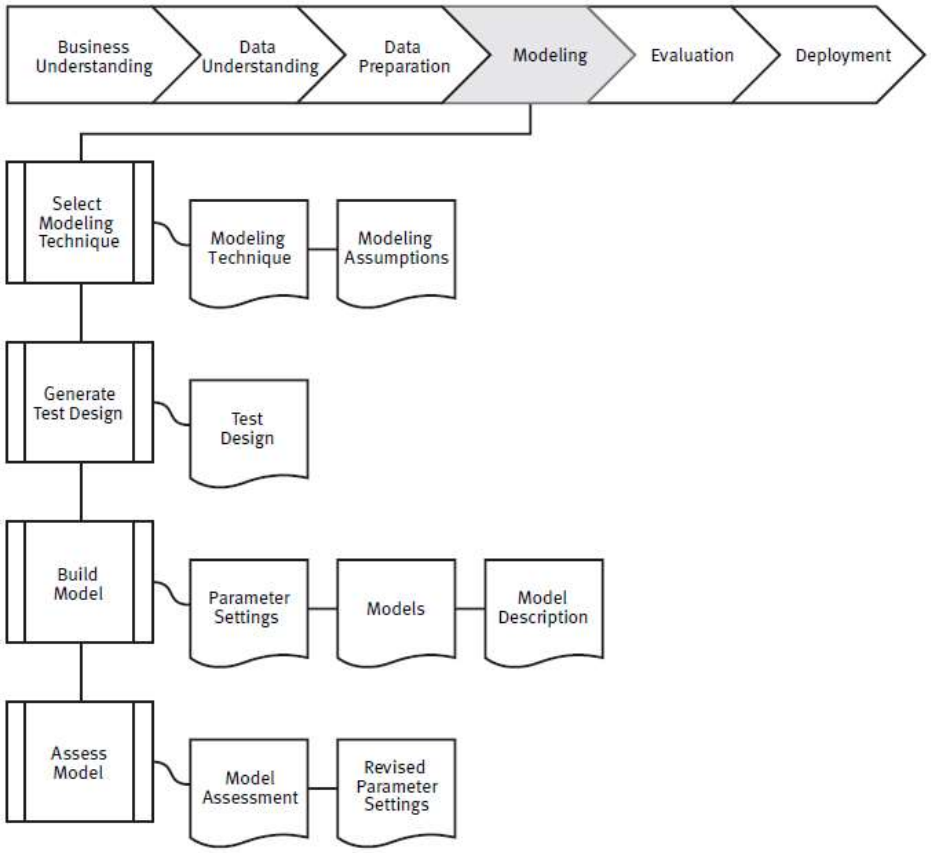
- Cubre toda actividad de construcción del conjunto de datos final que alimentará los modelos.
- Las tareas de preparación probablemente se repetirán varias veces.
- Selección de tablas, registros y atributos.
- Transformación de los datos originales.
- Limpieza de datos.



CRISP-DM

4. Modelado

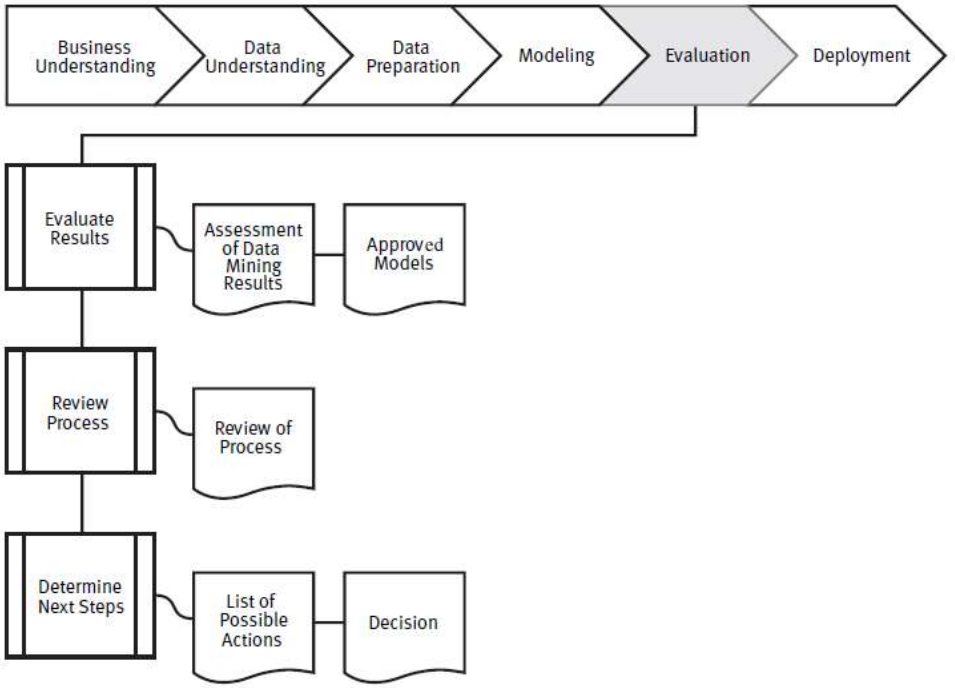
- Selección de técnicas y algoritmos de modelado que se van a aplicar.
- Calibración de los valores óptimos de los parámetros de los modelos.
- Usualmente existen varias técnicas que se pueden aplicar al mismo problema.
- Algunas técnicas tienen requerimientos muy específicos de los datos, por lo que es común tener que regresar a la fase anterior.
- Se estima la calidad de los modelos.



CRISP-DM

5. Evaluación

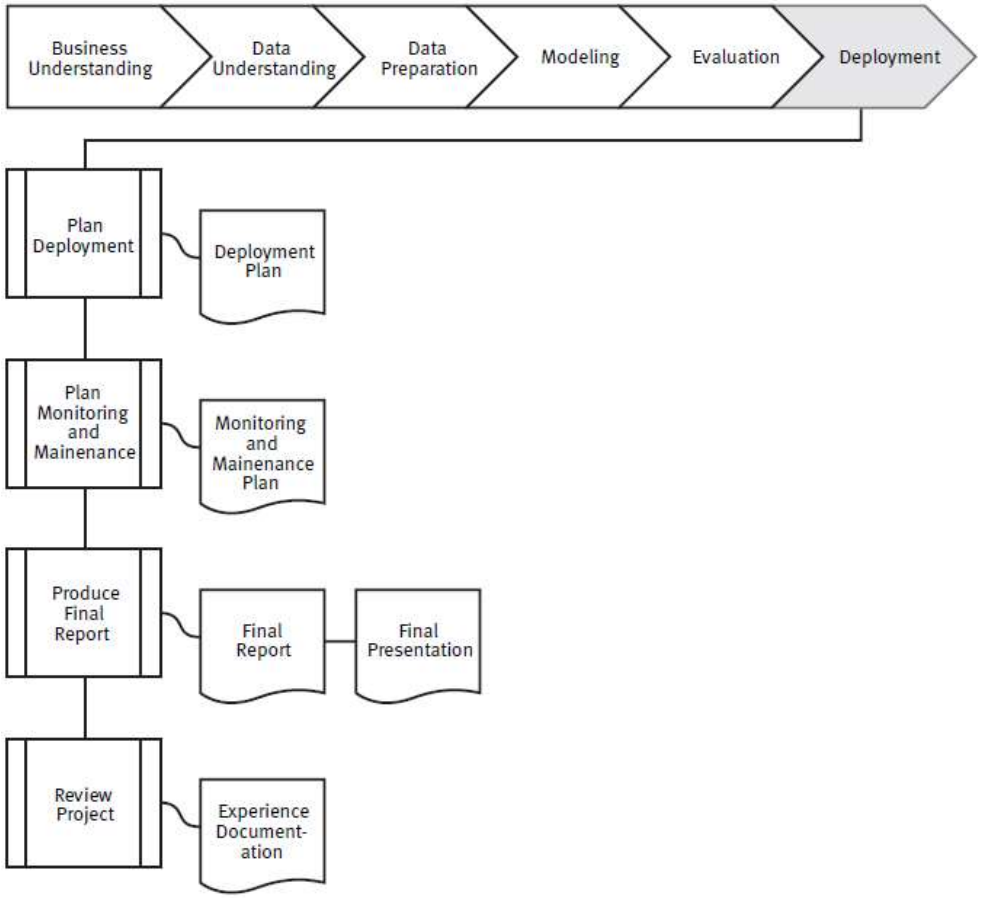
- En este punto, el(los) modelo(s) ya se han construido y validado desde el punto de vista de la analítica de datos.
- Antes de seguir al despliegue, es importante evaluarlo(s) a fondo y revisar los pasos tomados para crearlo(s), asegurándose que cumplen con los objetivos del negocio.
- Lo más importante es determinar si existe alguna particularidad del negocio que no se haya considerado.
- Al final de esta fase, se debe tomar una decisión de usar o no los resultados obtenidos.



CRISP-DM

6. Despliegue

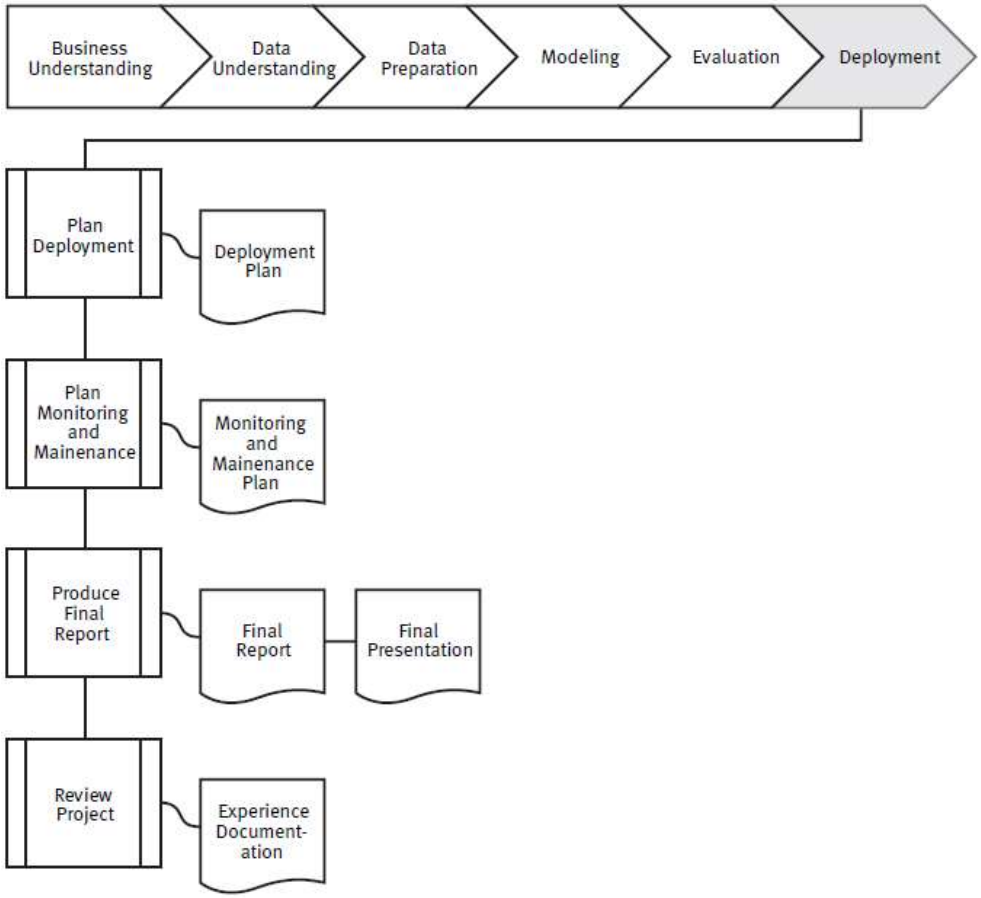
- El conocimiento alcanzado debe organizarse para presentarlo a los clientes.
- Integrar los modelos en los procesos de toma de decisión del negocio que presentaban los problemas/requerimientos iniciales.
- Dependiendo de los requerimientos, esta fase puede ser tan simple como generar un reporte estático, o tan compleja como implementar un proceso de analítica a lo largo de toda la organización.



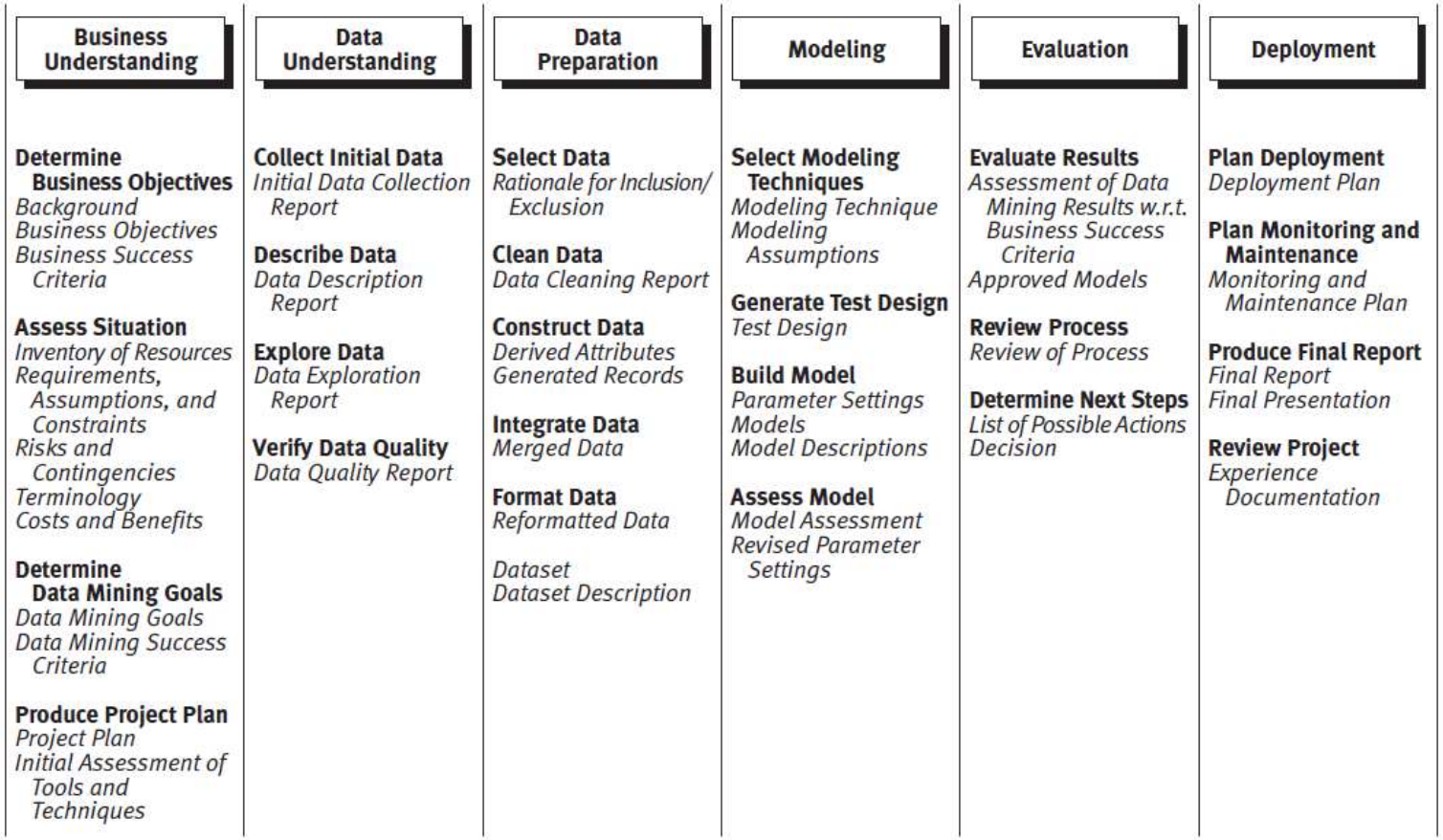
CRISP-DM

6. Despliegue (2)

- Es importante que el cliente comprenda las tareas necesarias para el despliegue, y que entienda cómo se utilizan.
- En muchos casos, es el cliente quien realiza el despliegue final.
- Al final, se regresa a la fase inicial de entendimiento del negocio.



CRISP-DM



Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Actividad

- Realice una lectura crítica de la guía paso a paso de CRISP-DM 1.0, escrita por Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS), Rüdiger Wirth (DaimlerChrysler), 2000. https://the-modeling-agency.com/crisp-dm.pdf

- **Nota 1:** recuerde que en el proyecto final puede seguir la metodología CRISP-DM.
- **Nota 2:** en la Maestría en Ciencia de Datos se explica también la metodología ASUM-DM *(Analytics Solutions Unified Method for Data Mining/Predictive Analytics*, IBM 2015). ASUM-DM es una extensión de CRISP-DM que tiene los mismos pasos en la minería de datos (desarrollo) más una parte operativa / de implementación. *ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf*