Programa de formación
MACHINE LEARNING AND DATA SCIENCE MLDS

Facultad de INGENIERÍA

UNIVERSIDAD NACIONAL DE COLOMBIA

Módulo 6

Metodologías de Desarrollo de Aplicaciones con Machine Learning

Unidad 1

Metodologías de Desarrollo de Machine Learning

Facultad de INGENIERÍA

UNIVERSIDAD NACIONAL DE COLOMBIA

Bienvenida

Jorge Eliécer Camargo Mendoza, PhD.

UNIVERSIDAD NACIONAL DE COLOMBIA

https://dis.unal.edu.co/~jecamargom/

jecamargom@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería
Universidad Nacional de Colombia
Sede Bogotá

Tabla de contenidos

Objetivos de aprendizaje

Unidad 1 - Metodologías de Desarrollo de Machine Learning

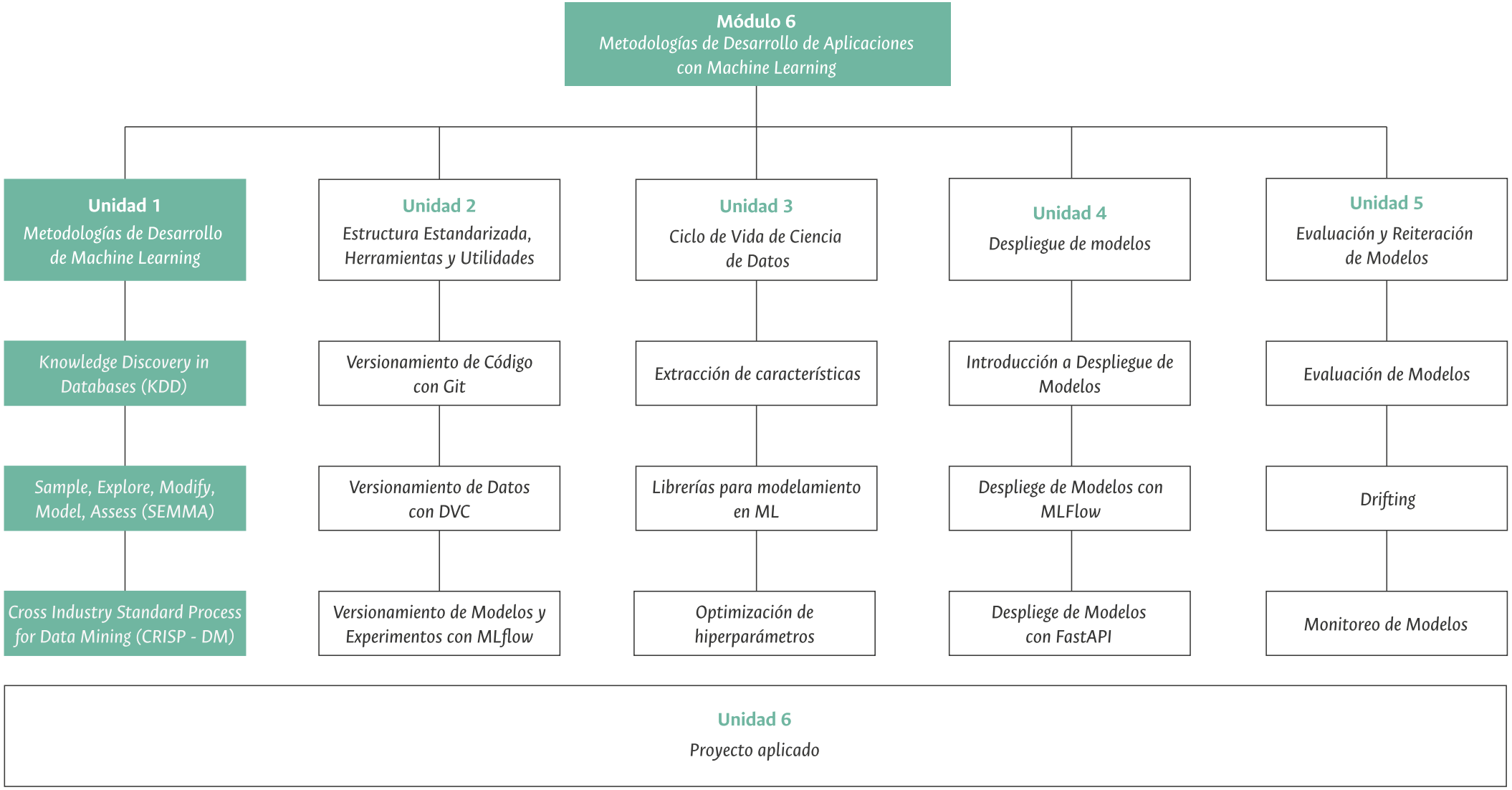
Al finalizar la unidad usted deberá ser capaz de:

Entender las distintas metodologías de desarrollo de aplicaciones de *Machine Learning*.

Identificar las diferencias entre metodologías de desarrollo de software y las de Ciencia de Datos.

Seleccionar la metodología de desarrollo más apropiada para Ciencia de Datos.

Mapa de contenidos de la unidad



Preguntas abiertas

- Y ahora que tengo el modelo de ML listo ¿Quién lo pone en producción?
- ¿Nace un proyecto de desarrollo de software?
- ¿En qué se debería desarrollar el software que utiliza el modelo de ML?
- ¿Cada cuánto debería evaluar nuevos modelos?
- ¿Cada cuánto tiempo debería entrenar el modelo de ML?
- ¿Qué pasa si tengo más datos?
- ¿Qué pasa mis datos ahora son de mejor calidad?
- ¿Qué pasa si tengo nuevos datos pero con más atributos?

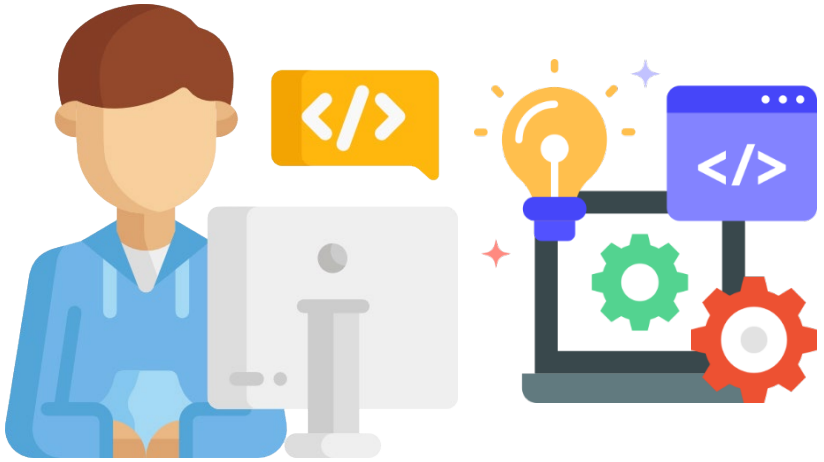
Ingeniería de Software

Artesanía

Ciencia

Arte

Arquitectura



Software

Ingeniería de software

Posibles definiciones

Zelkovitz, 1978

Estudio de los principios y metodologías para el desarrollo y mantenimiento de sistemas software.

Bohem, 1976

Aplicación práctica del conocimiento científico al diseño y construcción de programas de computadora y a la documentación asociada requerida para desarrollar, operar y mantenerlos

Trata del establecimiento de los principios y métodos de la ingeniería a fin de obtener software de modo rentable, que sea fiable y trabaje en máquinas reales

Bauer, 1973

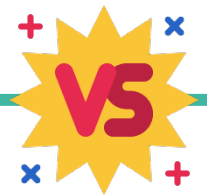
Es la aplicación de un enfoque sistemático, disciplinado y cuantificable al desarrollo, operación, y mantenimiento del software.

Standard Glossary of Software Engineering Terminology

Desarrollo de Software en Machine Learning

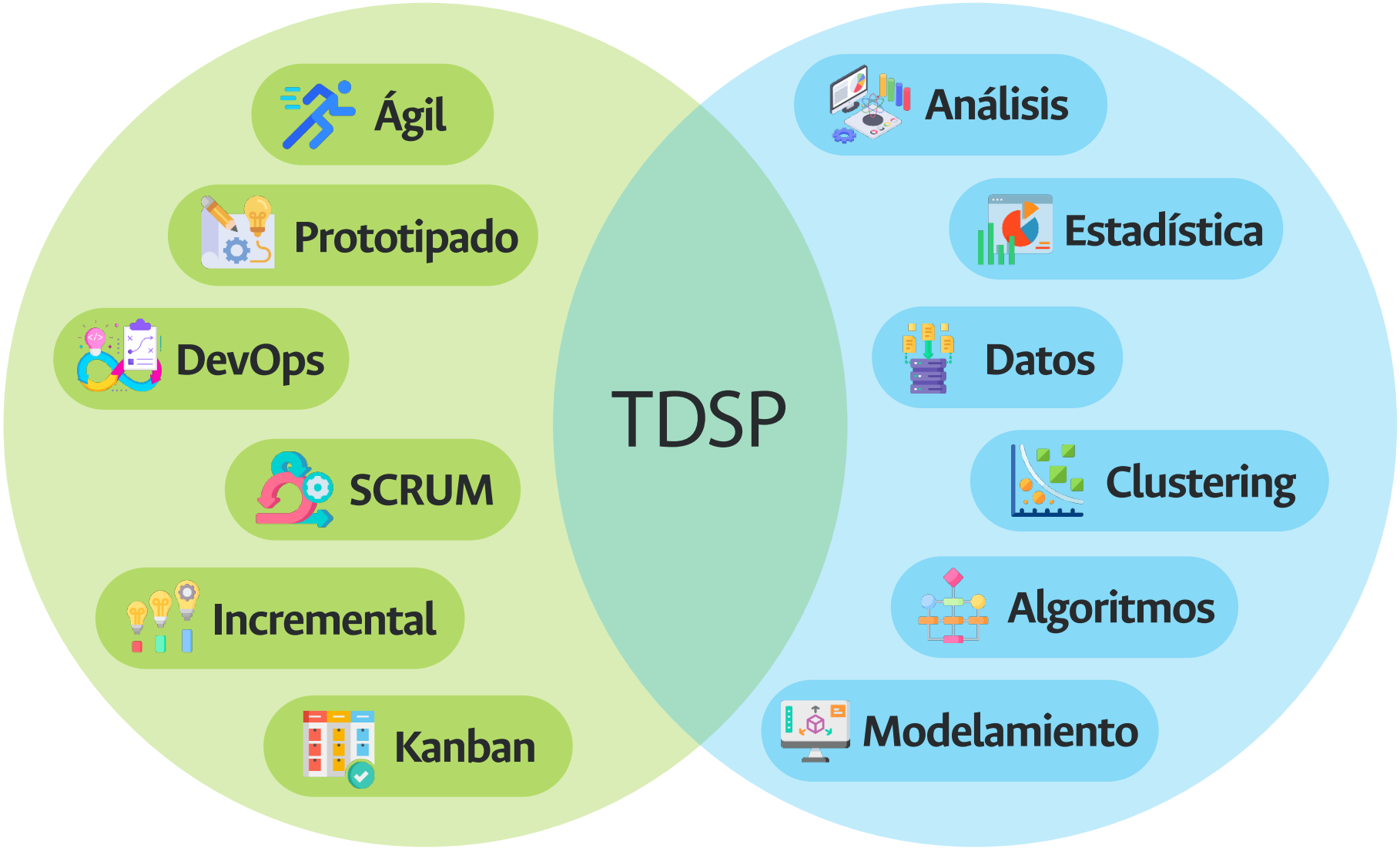
Proyecto sin Estructurar

```python
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

d = pd.read_csv('x.csv')

X = data.drop('c', axis=1)
y = data['c']

X1, xT, yT, Y1 = train_test_split(X, y, test_size=0.2, random_state=42)

t = DecisionTreeClassifier(random_state=42)
t.fit(X1, yT)

yp = t.predict(xT)

a = accuracy_score(Y1, yp)
print("P: {:.2f}%".format(accuracy * 100))
```

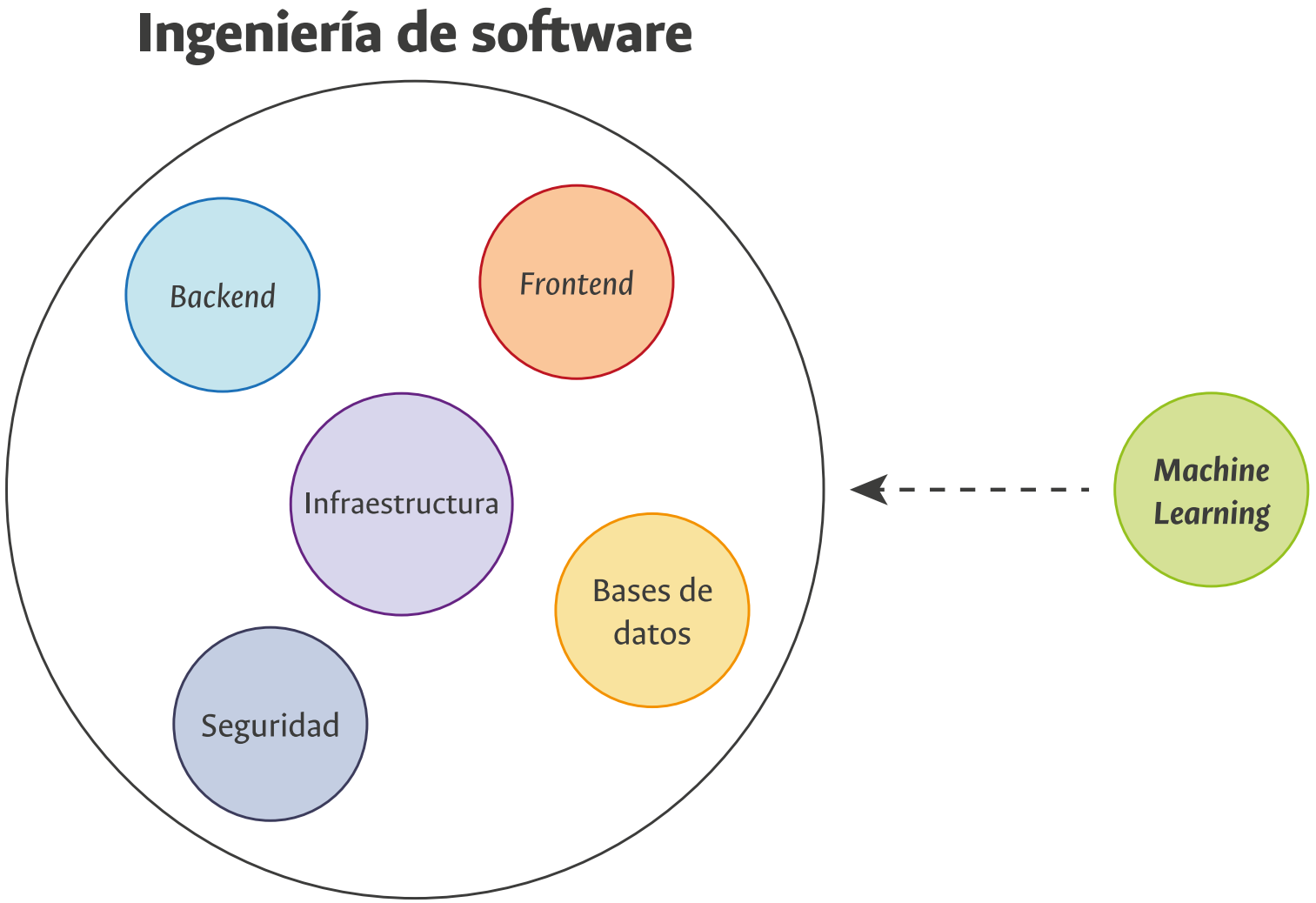


Proyecto Estructurado

Metodologías clásicas vs metodologías de *Machine Learning*



Machine Learning Operations (MLOps)



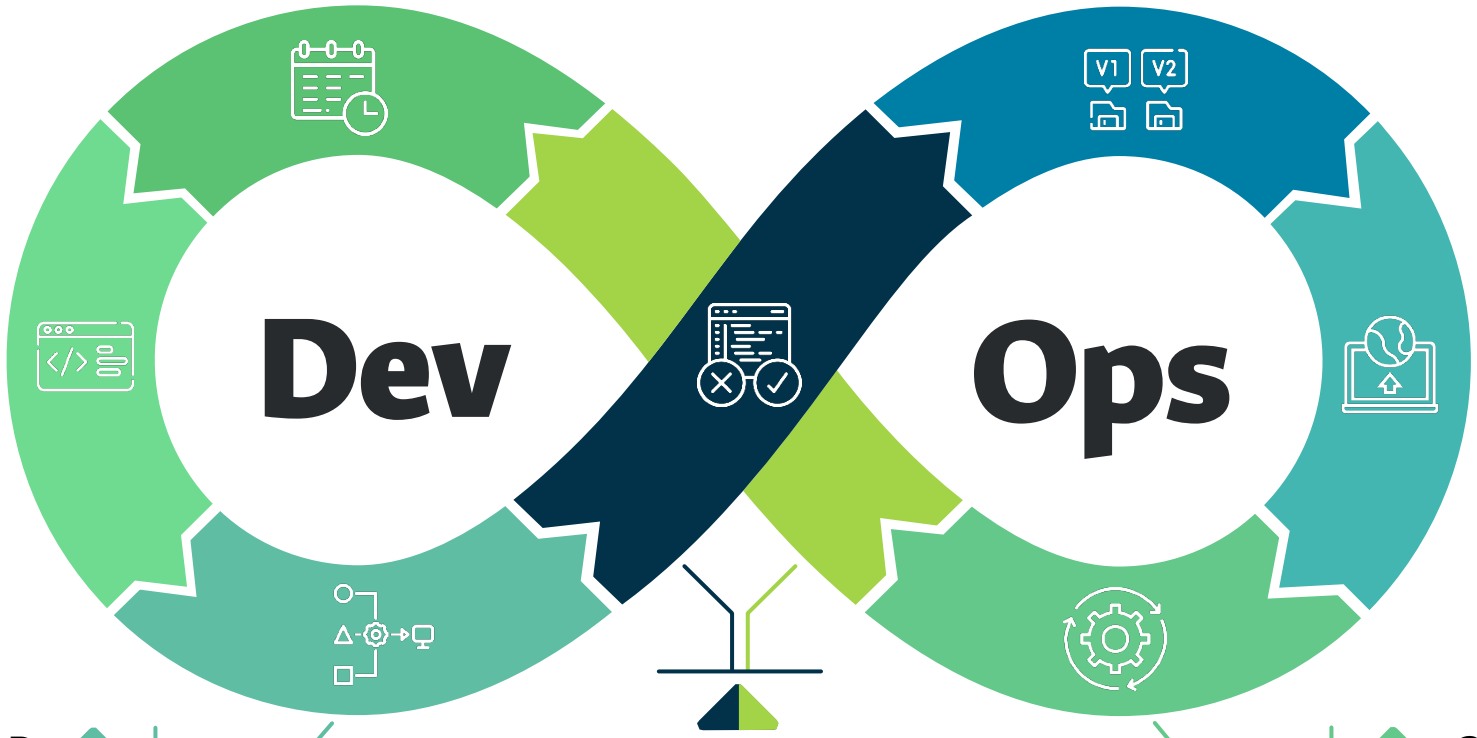
Machine Learning Operations (MLOps)

Ciclo de DevOps

PLANIFICAR
Pragma presente desde su concepción

CODIFICAR
Diseño adecuado y construcción de manera ágil

COMPILAR
Compilación óptima, verificando diseño, dependencias y versión.



PRUEBAS CONTINUAS Y MONITOREAR
Monitorear, analizar y medir el rendimiento de la aplicación y los recursos permanentes.

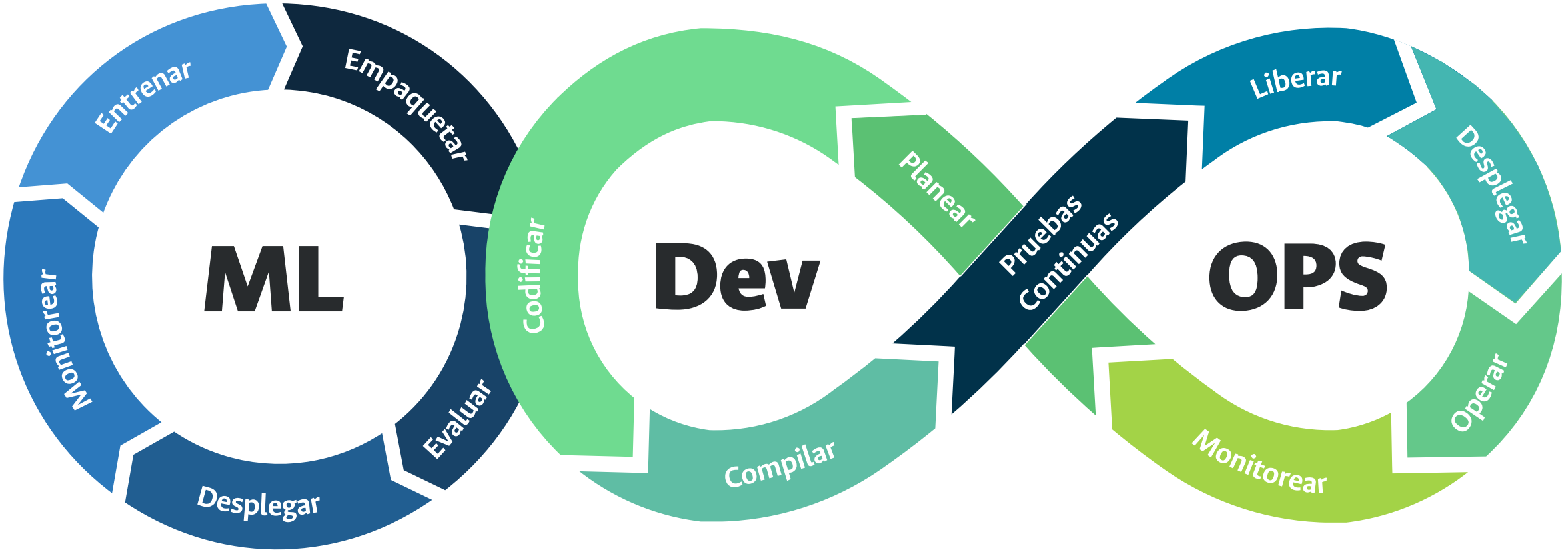
VERSIÓN
Producción de versiones que reflejan cambios de forma incremental.

DESPLEGAR
Despliegue continuo a los ambientes de producción.

OPERAR
Asegurar el funcionamiento adecuado de la aplicación después del despliegue.

Machine Learning Operations (MLOps)

MLOps



Tipos de metodologías en Ciencia de Datos

La mayoría de las metodologías plantean un proceso de manejo de datos, limpieza, modelamiento y evaluación.

- Existen distintas metodologías de desarrollo en ciencia de datos.
- Por lo general, este tipo de enfoques permiten definir una serie de tareas que posteriormente se pueden adaptar a una metodología de desarrollo de software.

1985

SEMMA

Sample, Explore, Modify, Model, Assess

1993

KDD

Knowledge Discovery in Databases

1996

CRISP – DM

Cross Industry Standard Process for Data Mining

2016

TDSP

Team Data Science Process

Tipos de metodologías en Ciencia de Datos

Comparación de Metodologías

Etapa	KDD	SEMMA	CRISP-DM	TDSP
1	Pre KDD	✕	Entendimiento del negocio	Entendimiento del negocio
2	Selección	Muestreo	Entendimiento de los datos	Adquisición y entendimiento
3	Preprocesamiento	Exploración	Entendimiento de los datos	Adquisición y entendimiento
4	Transformación	Modificación	Preparación de los datos	Adquisición y entendimiento
5	Minería	Modelamiento	Modelamiento	Modelamiento
6	Interpretación	Evaluación	Evaluación	Modelamiento
7	Post KDD	✕	Despliegue	Despliegue

SEMMA (Sample, Explore, Modify, Model, Assess)

Se trata de una lista de pasos secuenciales (cascada) que fue propuesta por el SAS Institute.

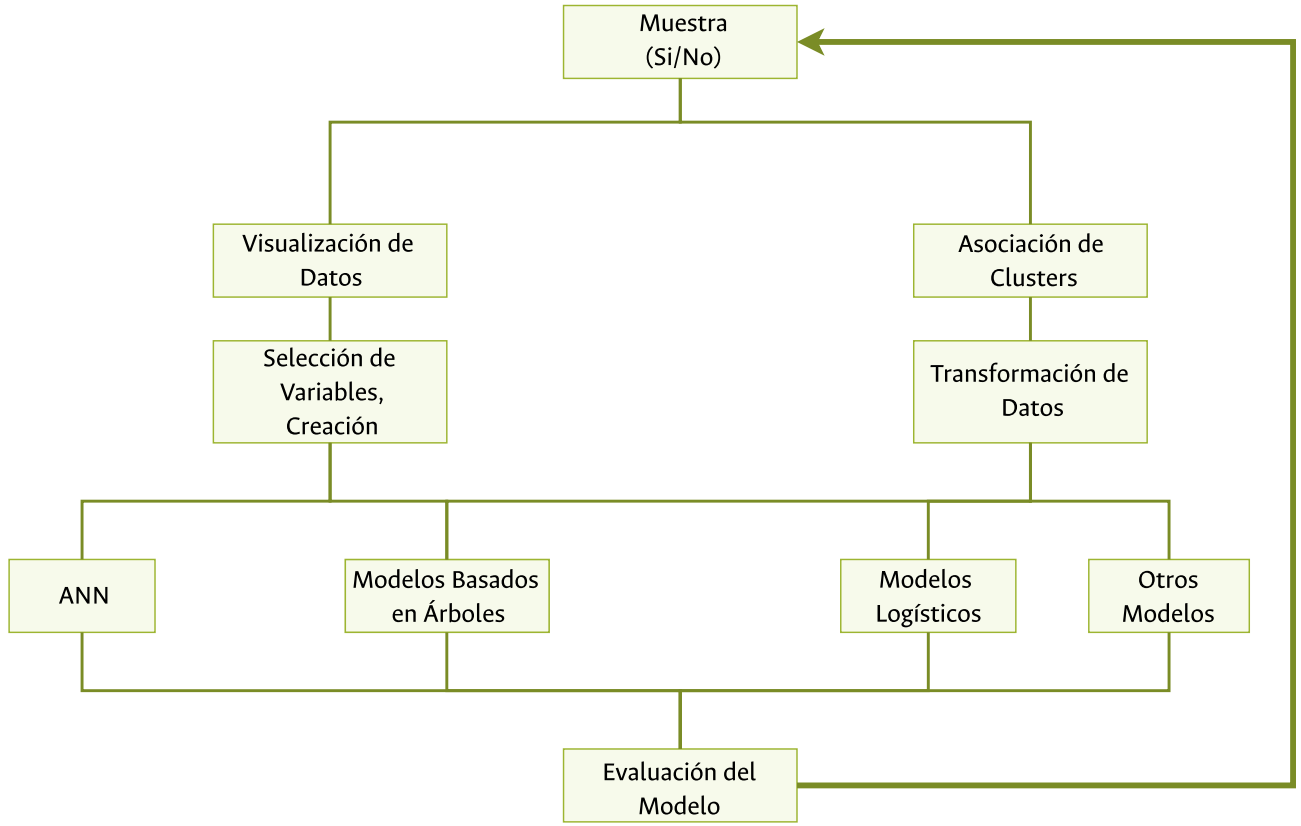
Sample (Muestrear)

Explore (Explorar)

Modify (Modificar)

Model (Modelar)

Assess (Evaluar)



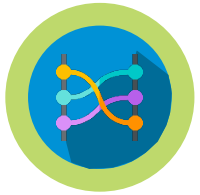
El planteamiento es similar a KDD, pero con un enfoque más estadístico.

SEMMA

Pasos

Sample (Muestrear)

Busca encontrar una muestra de volumen apropiado y la identificación de factores relevantes para el proceso.



Explore (Explorar)

Análisis univariado y multivariado se realiza para determinar relaciones entre los elementos y para determinar elementos faltantes en los datos. La visualización influye mucho en este proceso.

Modify (Modificar)

Tiene en cuenta lo encontrado en la exploración y busca extraer información y limpiar, preparando los datos para el modelamiento

Model (Modelar)

Es la aplicación de técnicas de minería de datos para generar un modelo.

Assess (Evaluar)

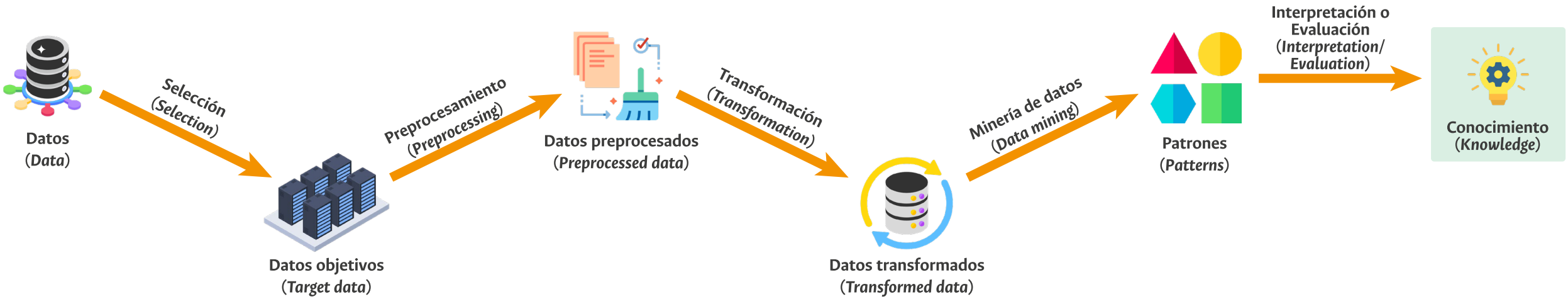
El modelo es evaluado con la finalidad de saber qué tan útil y confiable es.

KDD - Knowledge Discovery in Databases

Descubrimiento de Conocimiento en Bases de Datos (KDD)

El término Knowledge Discovery in Databases (KDD) fue usado por primera vez por Gregory Piatetsky-Shapiro en 1989.

Se trata de una metodología en cascada que consiste en un flujo bien definido basado en el análisis de datos.



KDD

Pasos

01 **Selection (Selección)** : Toma como insumo una base de datos (estructurada o no estructurada) y busca seleccionar los elementos (registros, variables o campos) relevantes para el análisis.

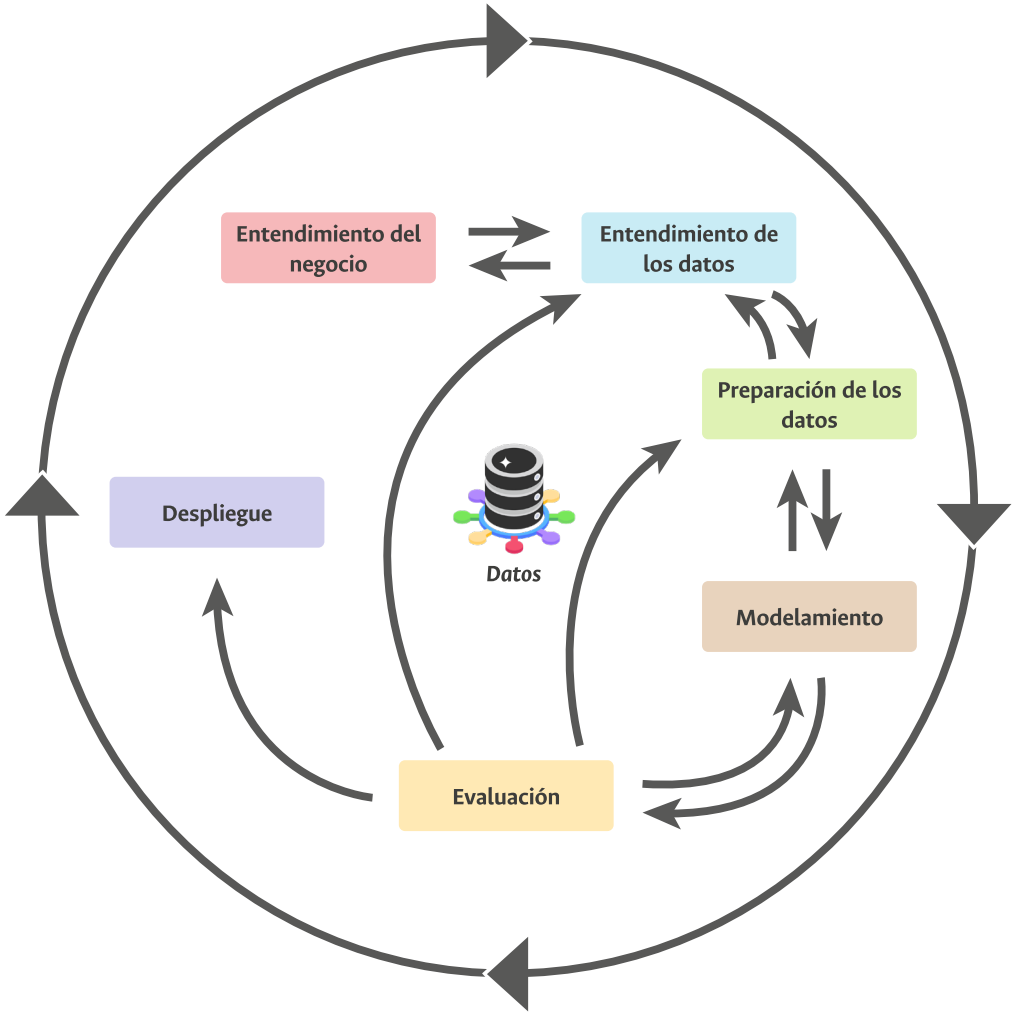
02 **Preprocessing (Preprocesamiento)** : Parte de los datos seleccionados, busca filtrar y transformar (en el dominio de los datos) los datos seleccionados.

03 **Transformation (Transformación)** : Es el proceso en el que los datos preprocesados se llevan a una representación numérica (matrices de observación y etiquetas).

04 **Data Mining (Minería de Datos)** : Toma como entrada los datos preprocesados y busca extraer patrones por medio de algún modelo estadístico o de Machine Learning.

05 **Interpretation and Evaluation (Interpretación y Evaluación)** : Parte de los patrones encontrados por los modelos para dar interpretatividad o evaluar el desempeño.

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



Proceso Estándar entre Industrias para la Minería de Datos

1. Se trata de un proceso cíclico que consiste en 6 fases que describen el ciclo de vida típico de ciencia de datos.

2. Se publicó en 1999 como estandarización de la minería de datos en la industria y desde entonces es la metodología más típica.

CRISP-DM

Fases

Fases de CRISP-DM

Business Understanding → **Entendimiento del Negocio** : Busca comprender los objetivos del proyecto, entendimiento de las necesidades del cliente, definición de Stakeholders y creación de un problema de minería de datos.

Data Understanding → **Entendimiento de los Datos** : En esta fase se comienzan a recolectar datos y se identifican: problemas de calidad, tipos de variables, subconjuntos de relevancia, interpretación de variables

Data Preparation → **Preparación de los Datos**: Consiste en todas las operaciones necesarias para construir el conjunto final de datos, esto incluye transformación, limpieza de datos y Data Wrangling.

Modeling → **Modelamiento** : Busca entrenar los modelos pertinentes al problema, clasificación, regresión, agrupamiento, reducción de dimensionalidad, entre otros.

Evaluation → **Evaluación** : Consiste en evaluar métricas (muchas veces referidas como key performance indicator) para determinar qué tan bien funciona el modelo.

Deployment → **Despliegue** : Proceso de despliegue del modelo, está relacionado al desarrollo de software donde se busca tener algún servicio, aplicación o librería que permita aplicar el modelo entrenado.

Team Data Science Process (TDSP) - Proceso de Ciencia de Datos en Equipo

Los componentes de TDSP tienen como objetivo resolver los retos del proceso DS.

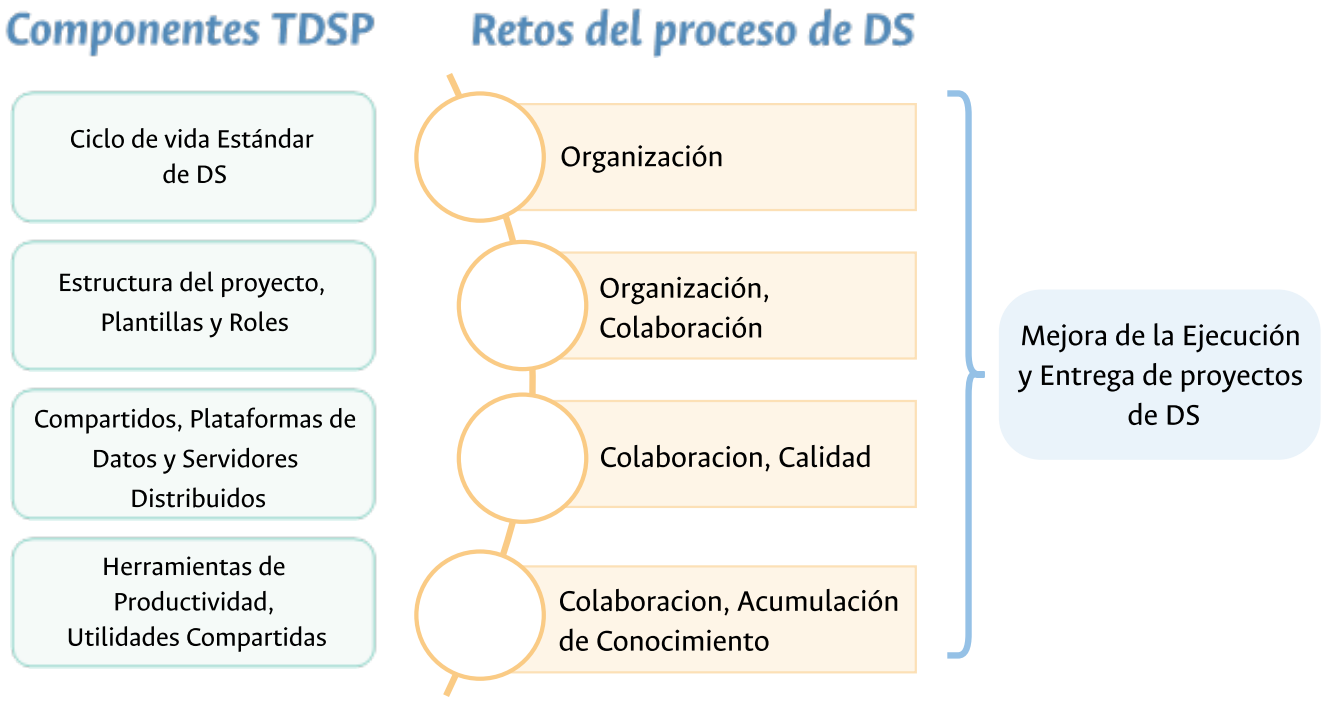


Diagrama conceptual de TDSP.

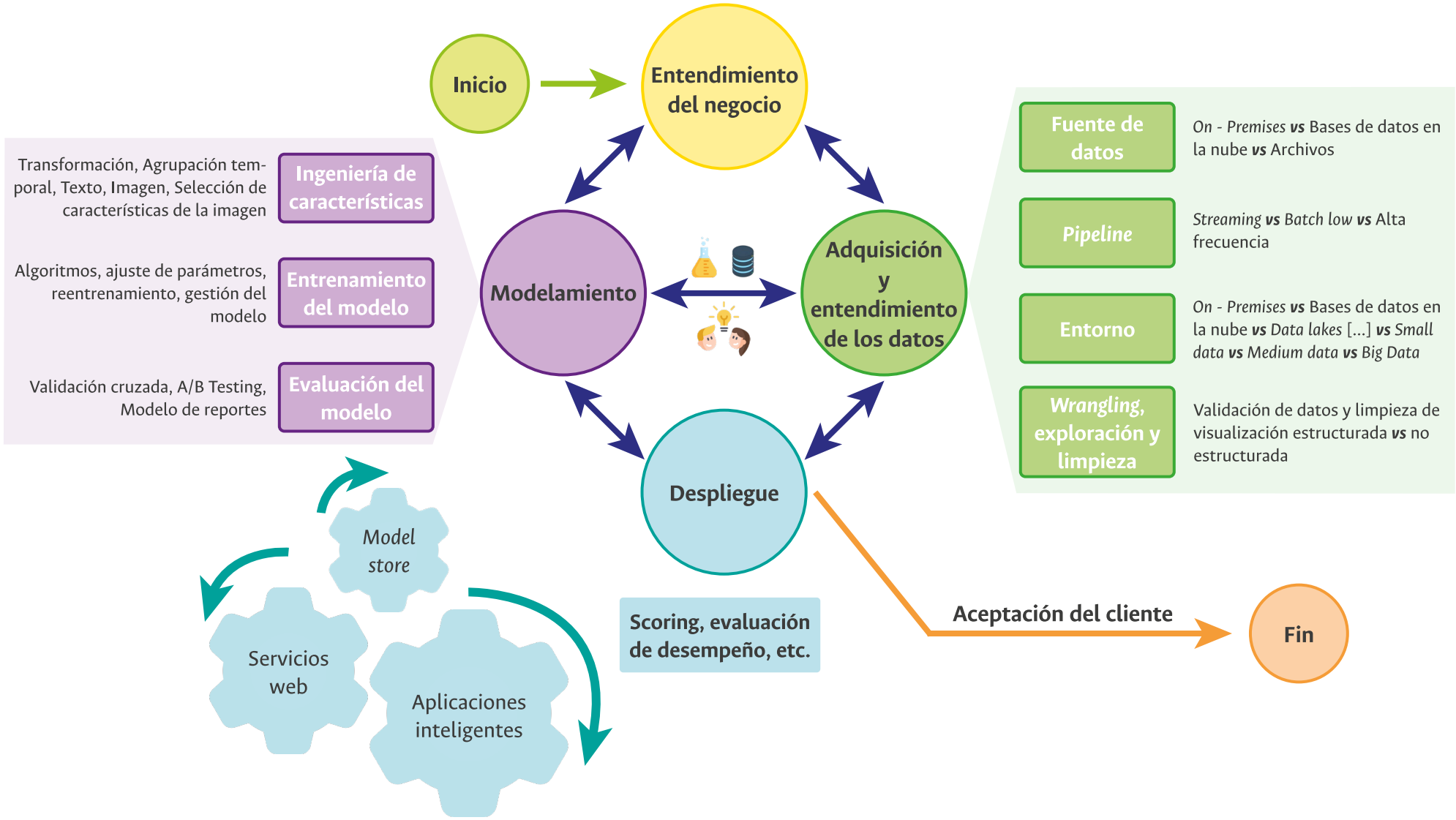
Se trata de una metodología que fue propuesta por Microsoft en el 2016.

Es una **metodología ágil** que busca estructurar el desarrollo de proyectos de Data Science en varios niveles :

- Ciclo de vida de ciencia de datos
- Infraestructura y recursos
- Estructuración del proyecto
- Herramientas y utilidades.

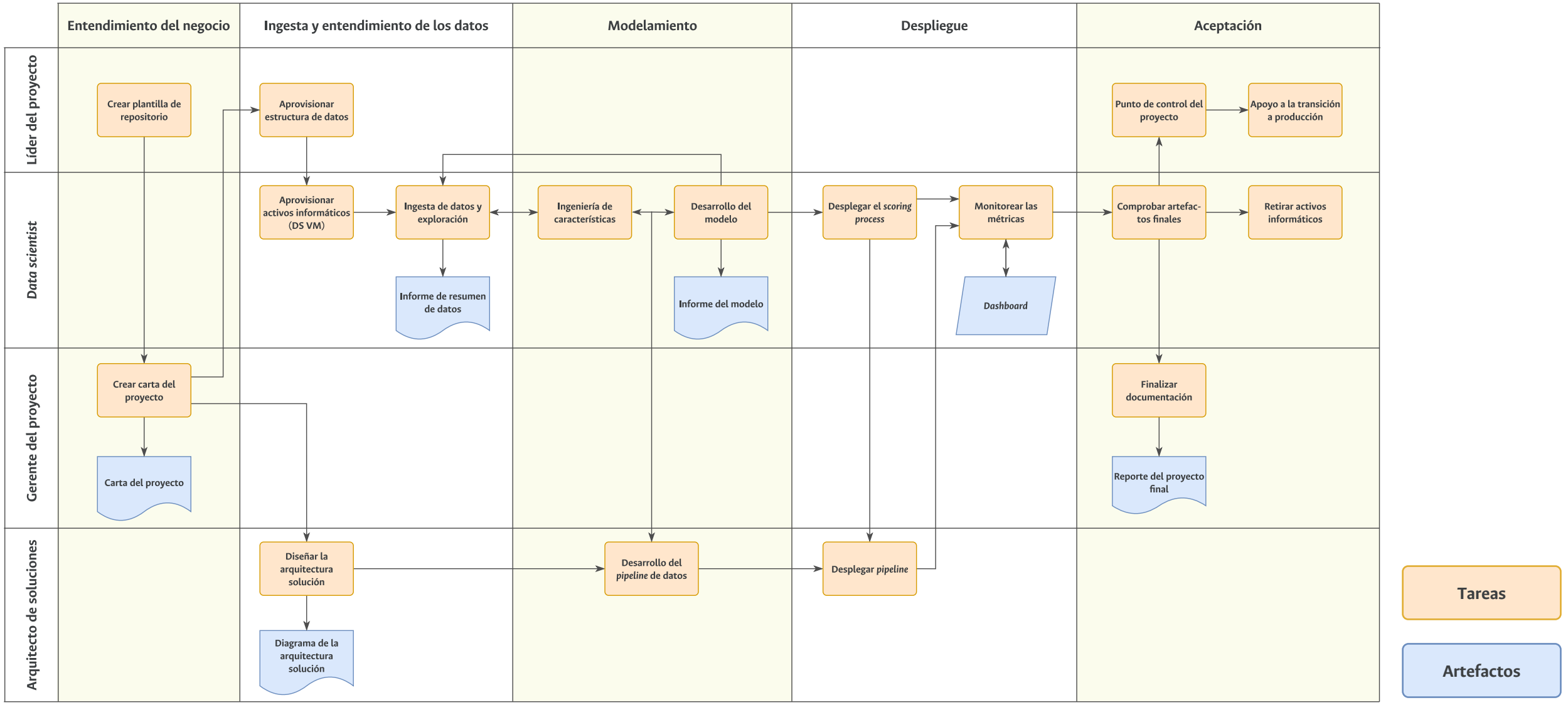
TDSP

Ciclo de Vida de la Ciencia de Datos



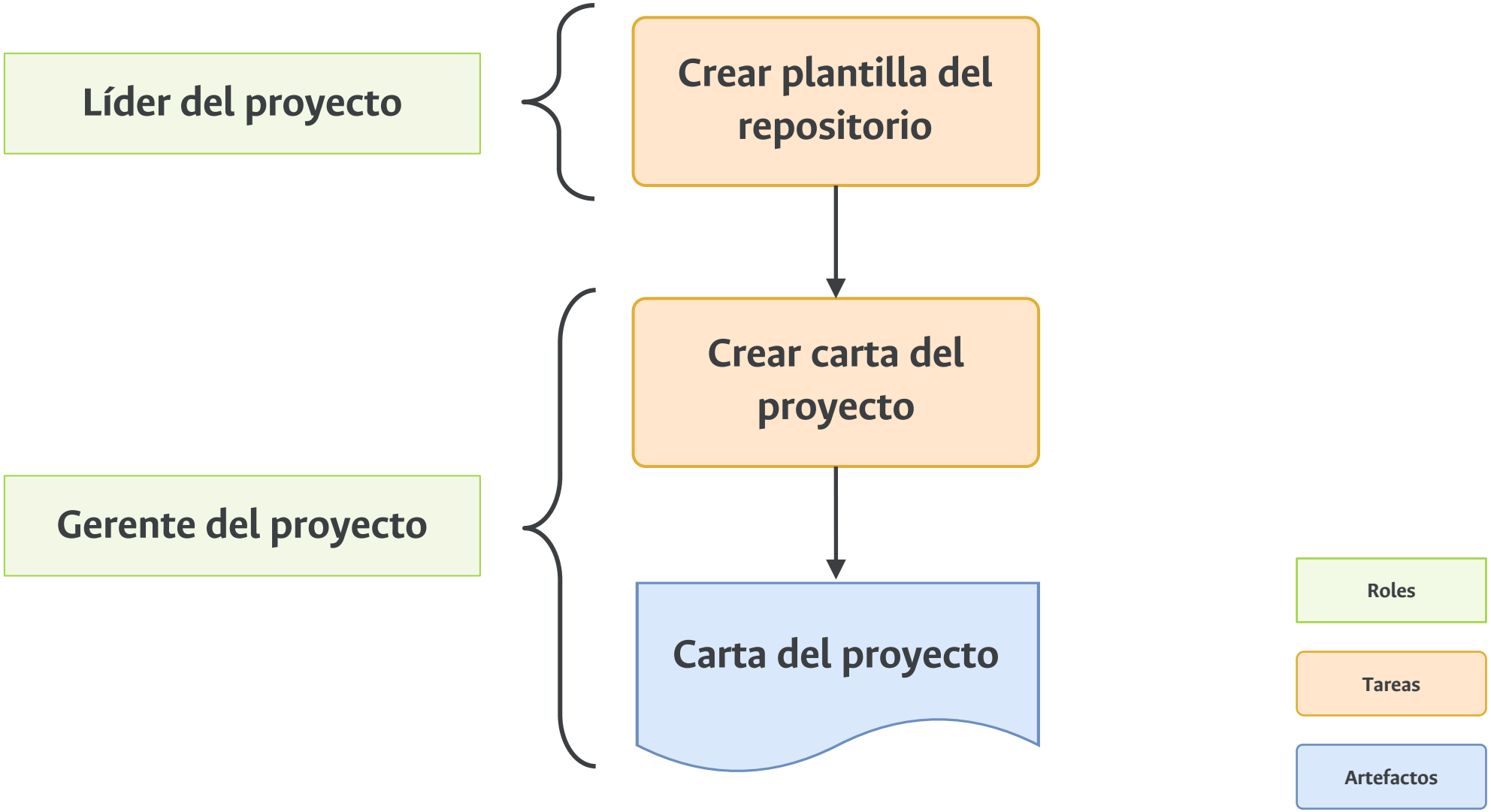
TDSP

Roles y Artefactos



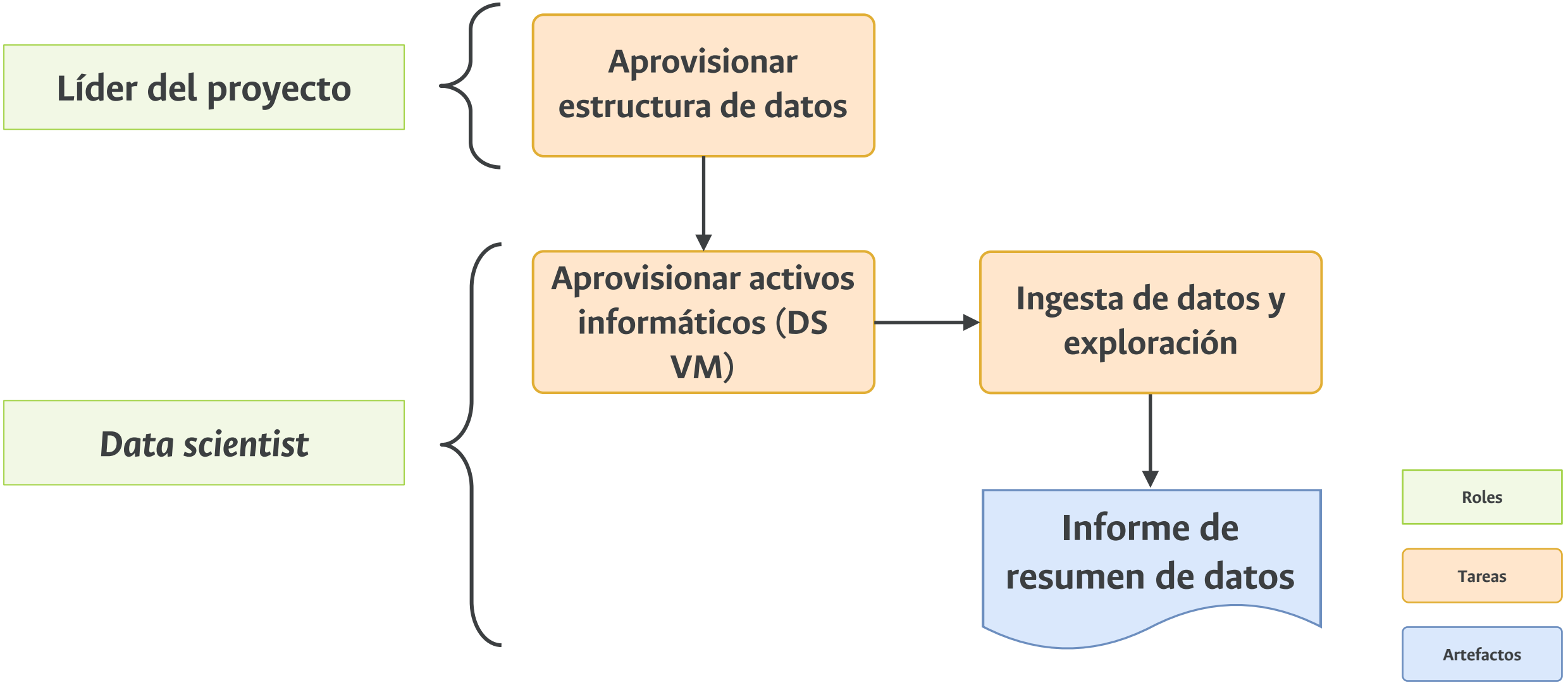
TDSP – Roles y Artefactos

Entendimiento del negocio



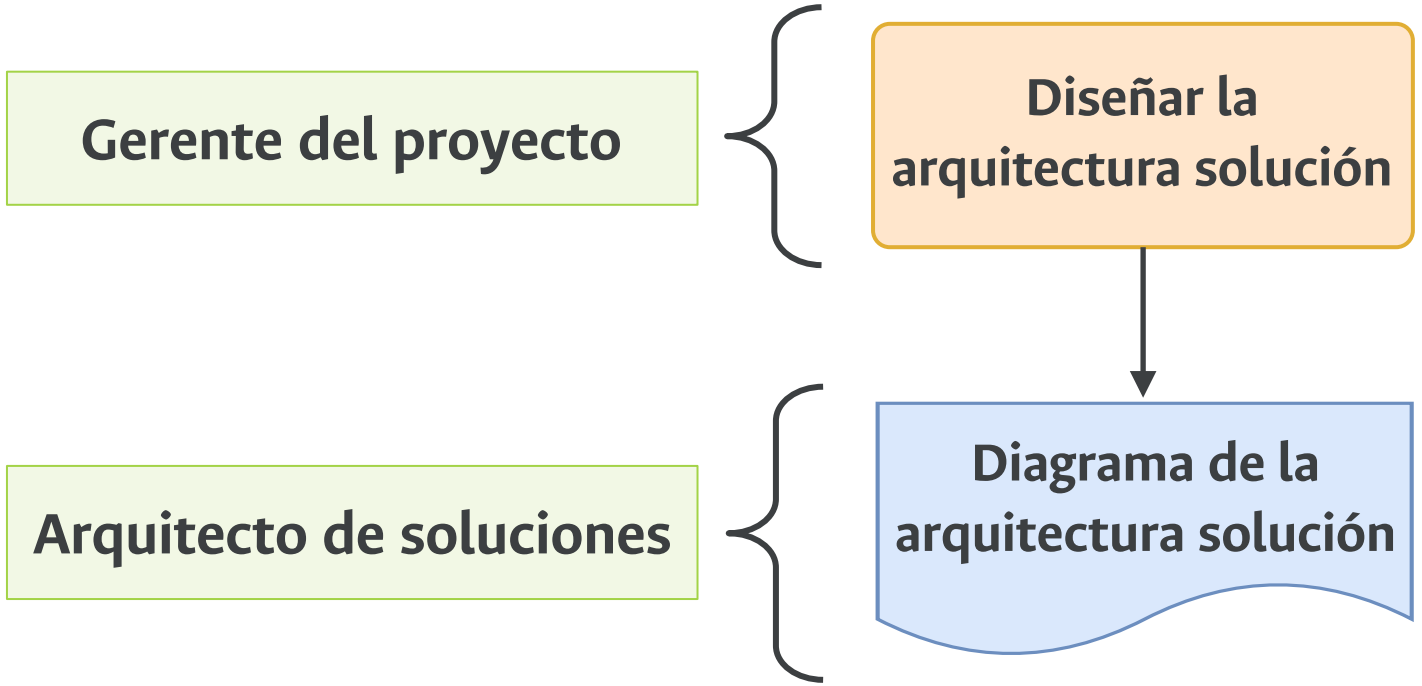
TDSP – Roles y Artefactos

Ingesta y entendimiento de los datos



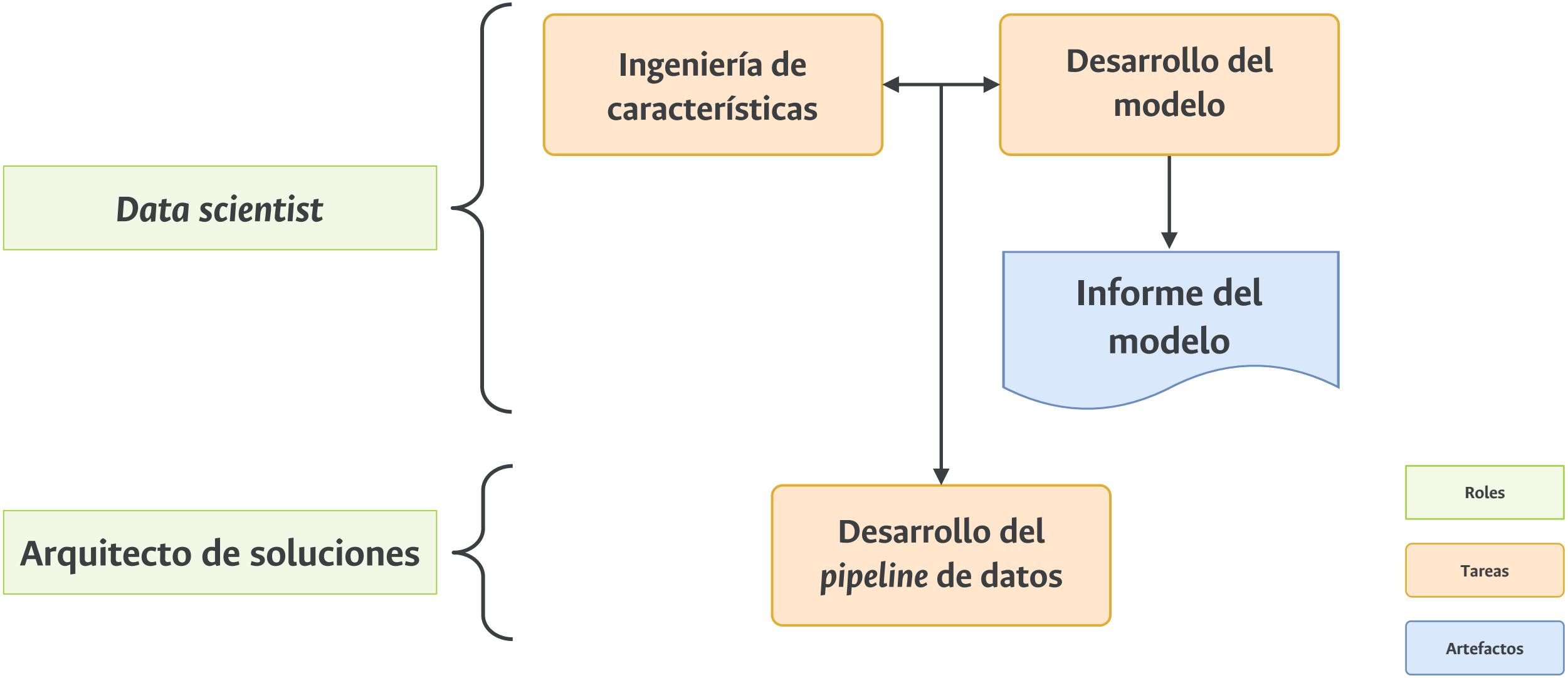
TDSP – Roles y Artefactos

Ingesta y entendimiento de los datos



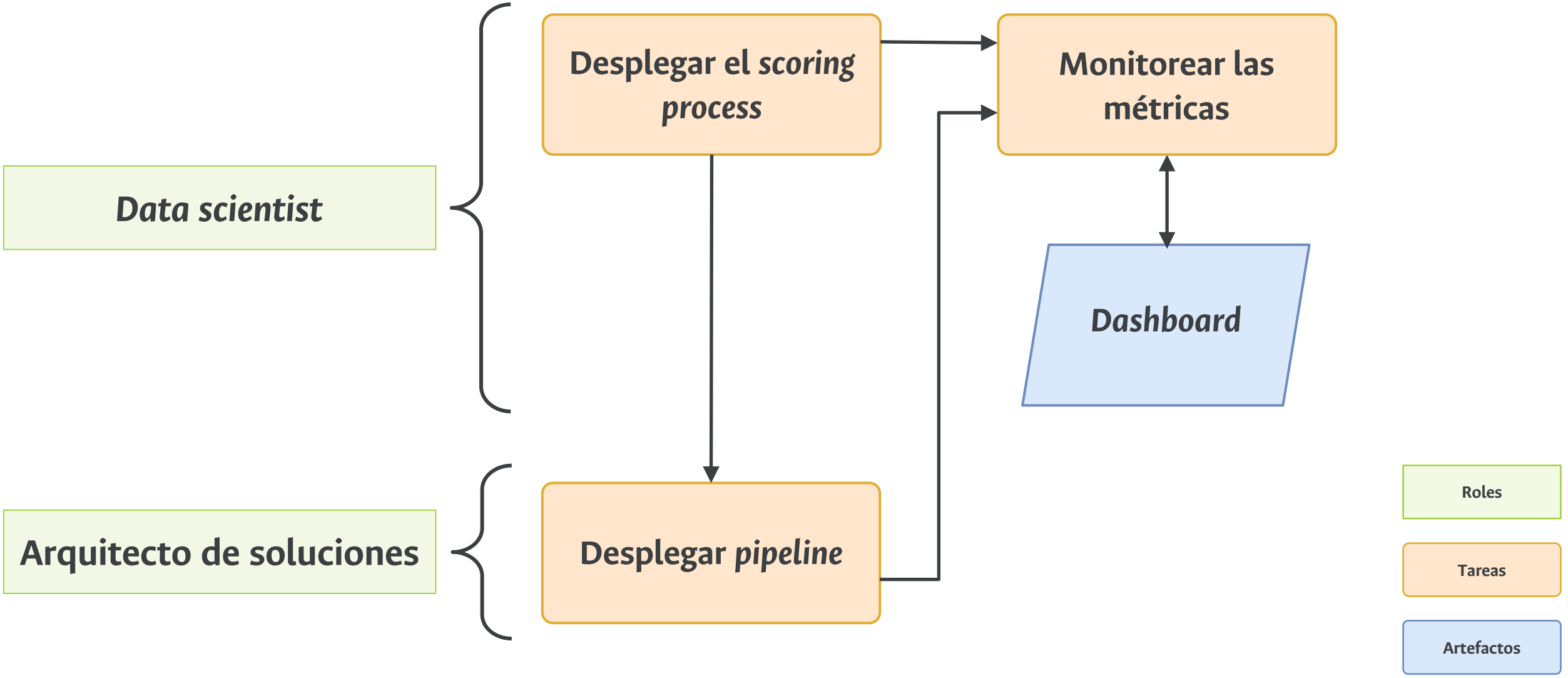
TDSP – Roles y Artefactos

Modelamiento



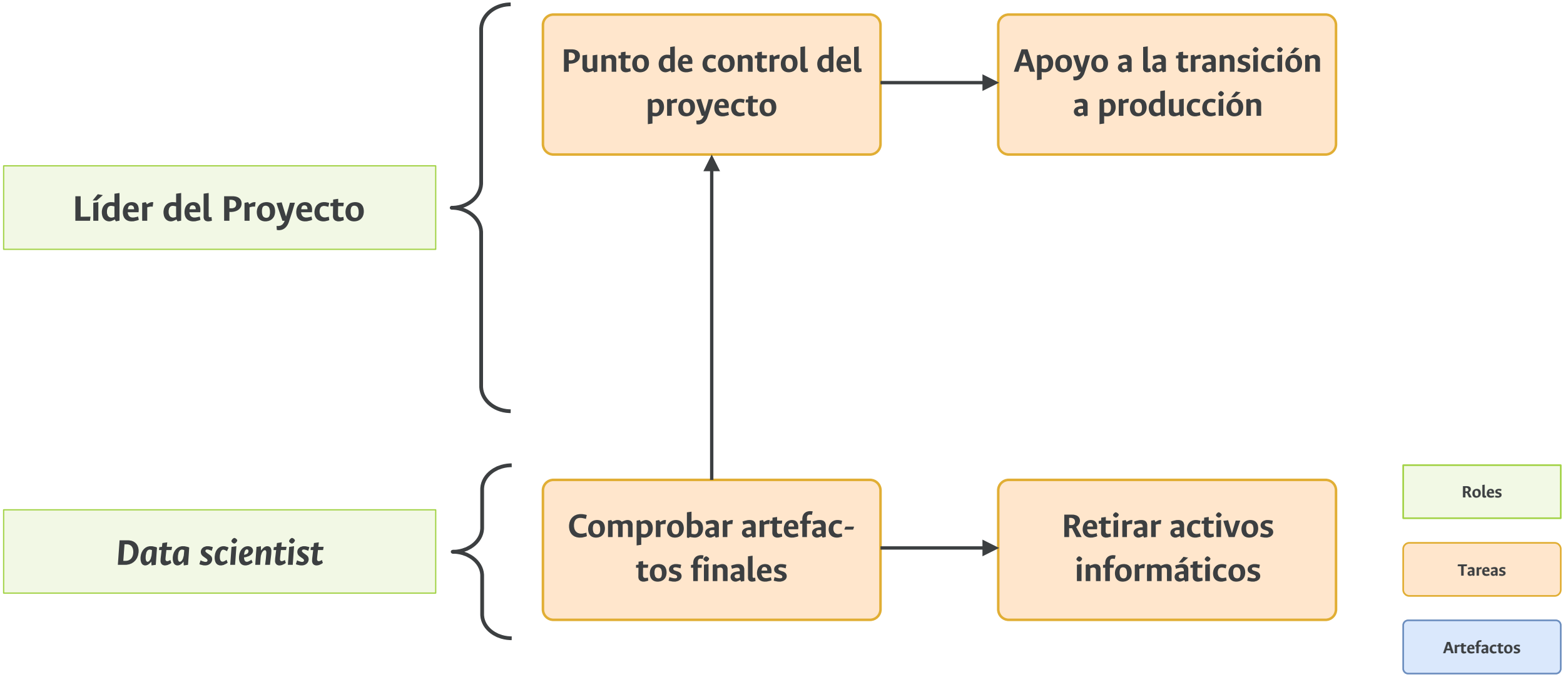
TDSP – Roles y Artefactos

Despliegue



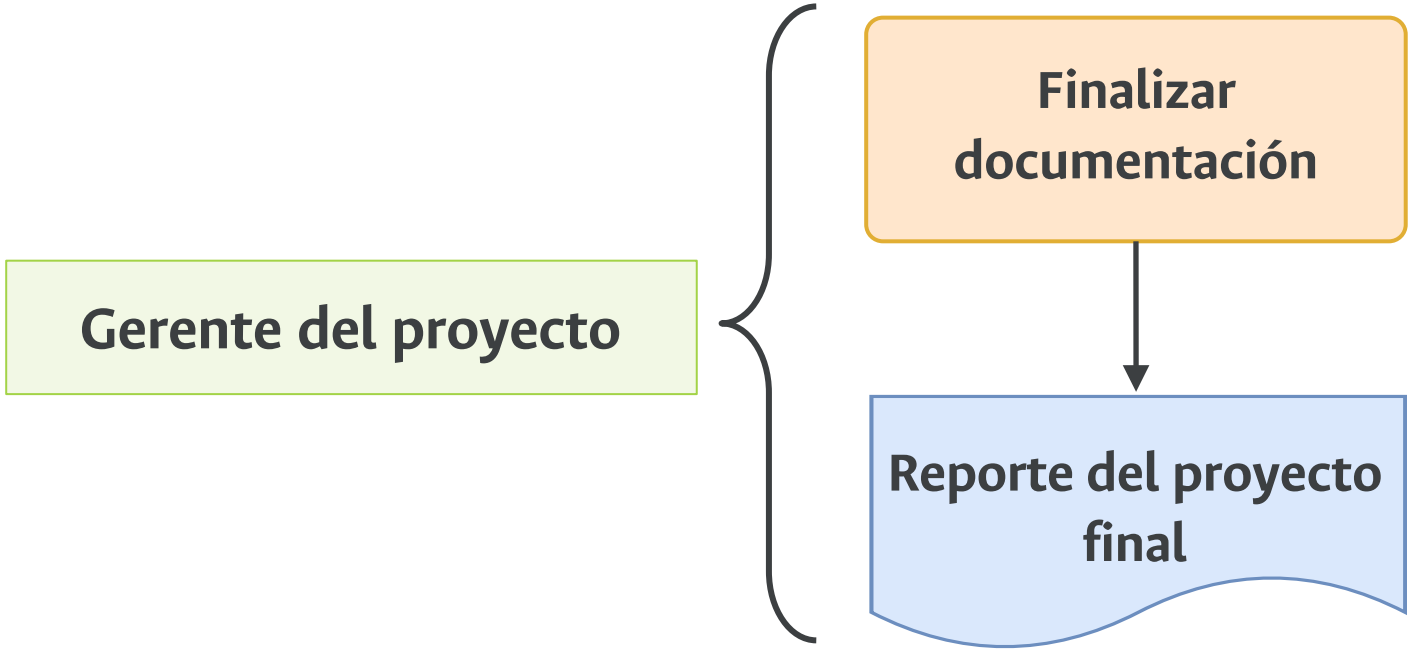
TDSP – Roles y Artefactos

Aceptación



TDSP – Roles y Artefactos

Aceptación



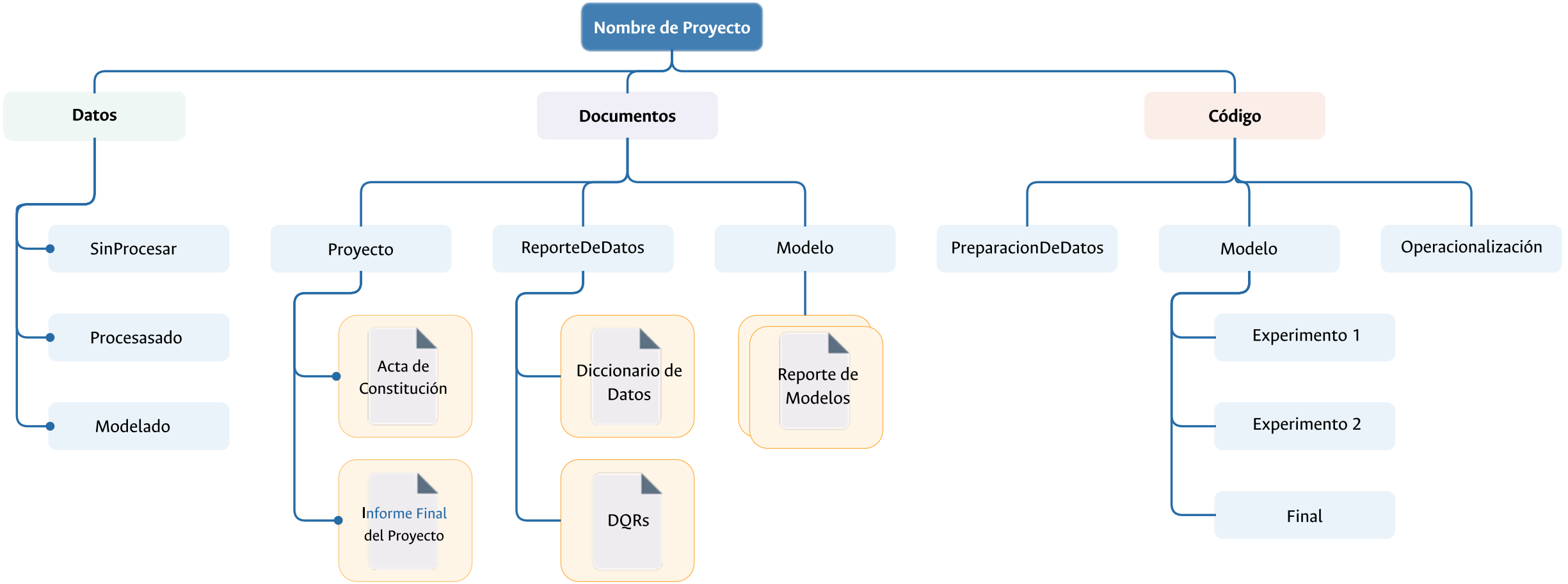
Roles

Tareas

Artefactos

TDSP

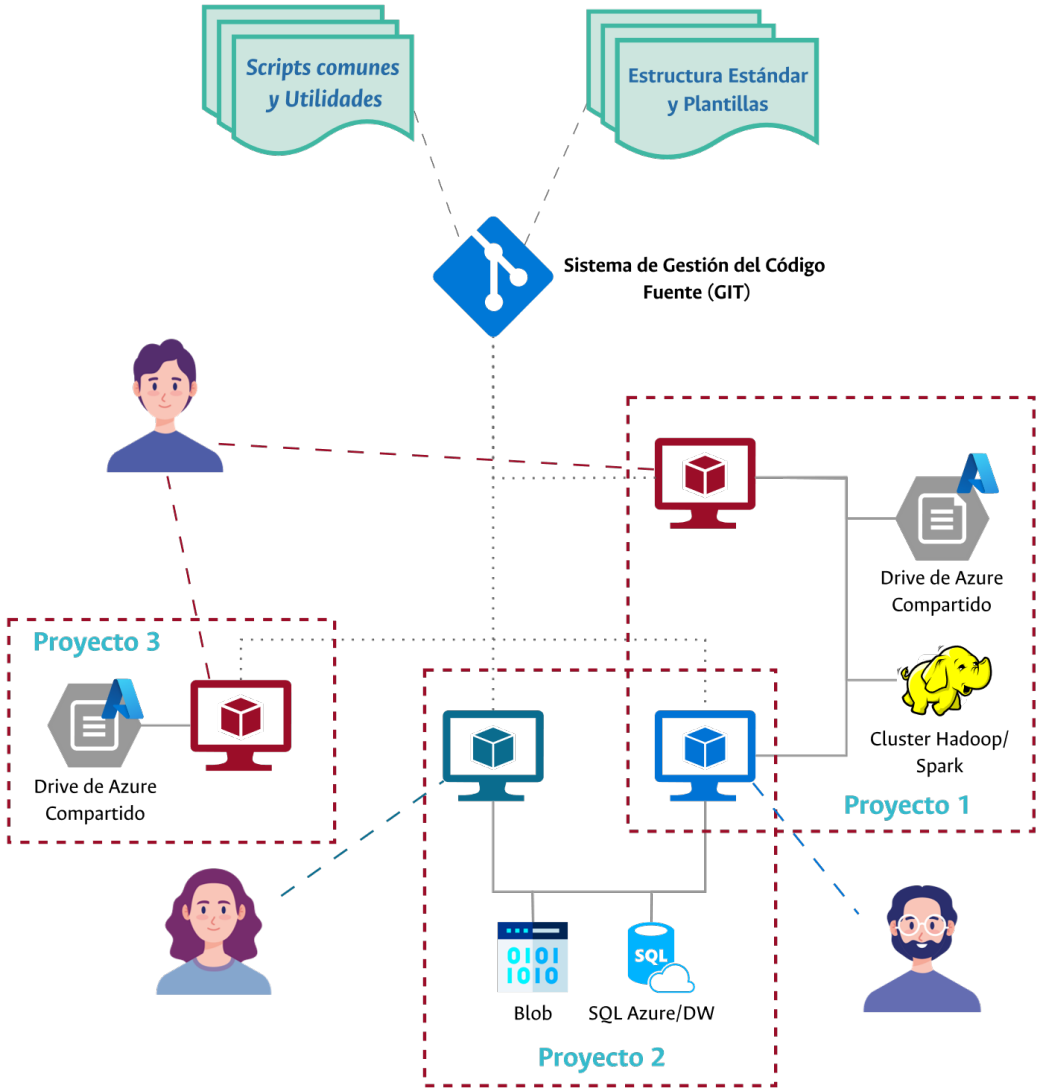
Estructura Estandarizada de Proyecto



Estructura Estandarizada de proyecto.

TDSP

Infraestructuras y Recursos



Infraestructura y Recursos para Ciencia de Datos.

Referencias

- R. S. Pressman y B. R. Maxim, Software Engineering: A practitioner's approach, 8th ed. New York, NY, USA: McGraw-Hill Education, 2015.
- Nogueira, C. Jones y Luqi, "Surfing the Edge of Chaos: Applications to Software Engineering," Monterey, CA, USA, 2000. Disponible: https://goo.gl/vXoAjK.
- I. Sommerville, Software Engineering, 10th ed. Essex, England: Pearson Education Limited, 2016.

- Modelo cascada y espiral, https://blog.comparasoftware.com/modelo-cascada-y-espiral/
- Scrum, https://desire.webs.uvigo.es/contenidos/scrum/
- University of Regina DBD. (s. f.). KDD Process/Overview. http://www2.cs.uregina.ca/%7Edbd/cs831/notes/kdd/1_kdd.html
- Hotz, N. (2023, 31 enero). What is SEMMA? Data Science Process Alliance. https://www.datascience-pm.com/semma/
- Hotz, N. (2023a, enero 19). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/
- M. (s. f.). What is the Team Data Science Process? - Azure Architecture Center. Microsoft Learn. https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview

Derechos de imágenes

- Flaticon. (s.f.). Relations icon. [Icono]. https://www.flaticon.com/free-icon/relations_1807303
- Flaticon. (s.f.). Broom icon. [Icono]. https://www.flaticon.com/free-icon/broom_2954880
- Flaticon. (s.f.). Data mining icon. [Icono]. https://www.flaticon.com/free-icon/data-mining_1991040
- Flaticon. (s.f.). Planning icon. [Icono]. https://www.flaticon.com/free-icon/planning_1322237
- Flaticon. (s.f.). Code icon. [Icono]. https://www.flaticon.com/free-icon/code_2920242
- Flaticon. (s.f.). Compile icon. [Icono]. https://www.flaticon.com/free-icon/compile_8185151
- Flaticon. (s.f.). Testing icon. [Icono]. https://www.flaticon.com/free-icon/testing_6403594
- Flaticon. (s.f.). Version icon. [Icono]. https://www.flaticon.com/free-icon/version_8083300
- Flaticon. (s.f.). Deployment icon. [Icono]. https://www.flaticon.com/free-icon/deployment_1508763
- Flaticon. (s.f.). Cogwheel icon. [Icono]. https://www.flaticon.com/free-icon/cogwheel_1988036
- Flaticon. (s.f.). Data server icon. [Icono]. https://www.flaticon.com/free-icon/data-server_2717155
- Flaticon. (s.f.). Shapes icon. [Icono]. https://www.flaticon.com/free-icon/shapes_1215843
- Flaticon. (s.f.). Update icon. [Icono]. https://www.flaticon.com/free-icon/update_9517843
- Flaticon. (s.f.). Craftswoman free icon. [Icono]. https://www.flaticon.com/free-icon/craftswoman_7879067
- Flaticon. (s.f.). Pottery free icon. [Icono]. https://www.flaticon.com/free-icon/pottery_5903201
- Flaticon. (s.f). Artist free icon. [Icono]. https://www.flaticon.com/free-icon/artist_3271406

Derechos de imágenes

- Flaticon. (s.f.). Painting free icon. [Icono]. https://www.flaticon.com/free-icon/painting_3370596
- Flaticon. (s.f.). Code free icon. [Icono]. https://www.flaticon.com/free-icon/code_2920242
- Flaticon. (s.f.). Social Science free icon. [Icono]. https://www.flaticon.com/free-icon/social-science_4459306
- Flaticon. (s.f.). Architect free icon. [Icono]. https://www.flaticon.com/free-icon/architect_5757029
- Flaticon. (s.f.). Building free icon. [Icono]. https://www.flaticon.com/free-icon/building_717940
- Flaticon. (s.f.). Programmer free icon. [Icono]. https://www.flaticon.com/free-icon/programmer_644658
- Flaticon. (s.f.). Software free icon. [Icono]. https://www.flaticon.com/free-icon/software_3950815
- Flaticon. (s.f.). File free icon. [Icono]. https://www.flaticon.com/free-icon/file_3155758
- Flaticon. (s.f.). Man free icon. [Icono]. https://www.flaticon.com/free-icon/man_4202843
- Flaticon. (s.f.). Girl free icon. [Icono]. https://www.flaticon.com/free-icon/girl_4202836
- Flaticon. (s.f.). Boy free icon. [Icono]. https://www.flaticon.com/free-icon/boy_4202831
- Flaticon. (s.f.). Binary free icon. [Icono]. https://www.flaticon.com/free-icon/binary_2592364
- Wikimedia. (2021, 3 agosto). Logo of Microsoft Azure. [Logo]. https://upload.wikimedia.org/wikipedia/commons/f/fa/Microsoft_Azure.svg
- WORLDVECTORLOGO. (s.f.). Hadoop vector. [Logo]. https://cdn.worldvectorlogo.com/logos/hadoop.svg

Despedida

¡Gracias por su atención!

Jorge Eliécer Camargo Mendoza, PhD.

https://dis.unal.edu.co/~jecamargom/

jecamargom@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería
Universidad Nacional de Colombia
Sede Bogotá

UNIVERSIDAD NACIONAL DE COLOMBIA

Créditos

Facultad de
INGENIERÍA

Profesor
Jorge Eliécer Camargo Mendoza, PhD

Asistente docente
Juan Sebastián Lara Ramírez

Coordinador de virtualización
Edder Hernández Forero

Diagramadores PPT
Mario Andrés Rodríguez Triana
Rosa Alejandra Superlano Esquibel

Diseño gráfico
Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

2024