DataOps como metodología de desarrollo en proyectos de Aprendizaje de Máquinas

David Fernando Rivera Olarte

Candidato a Magister en Ingeniería – Ingeniería de Sistemas

dfriveraol@unal.edu.co

Área Curricular de Sistemas e Informática
Facultad de Minas
Sede Medellín
2020

UNIVERSIDAD NACIONAL DE COLOMBIA

Agenda

- Introducción
- Comparativo de Metodologías
- Aproximación Metodológica
- Conclusiones

Agenda

- Introducción
- Comparativo de Metodologías
- Aproximación Metodológica
- Conclusiones

Introducción

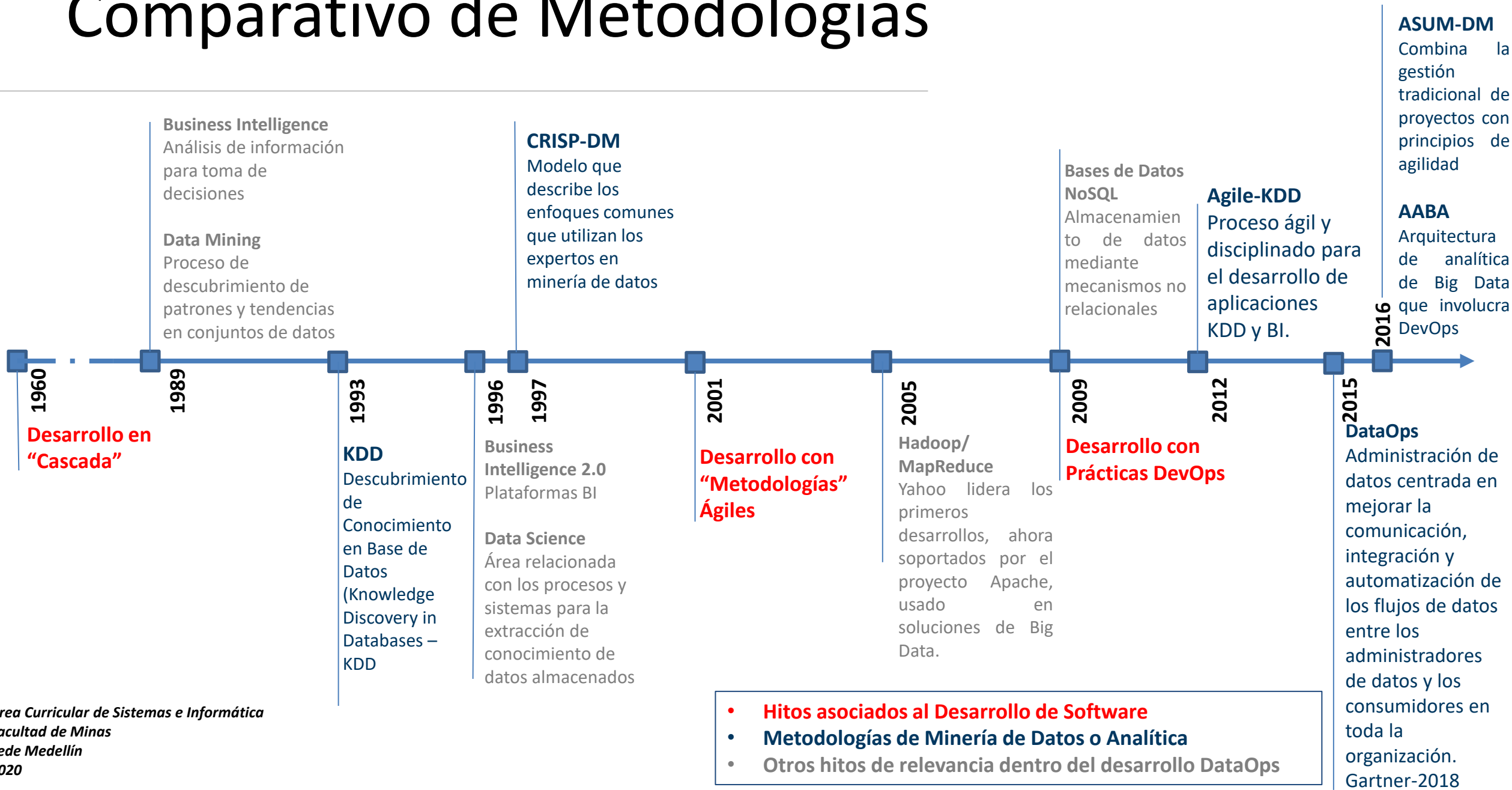
Por que las metodologías de desarrollo de SW no funcionan para ML.

- No hay momento “ágil”.
- Proyectos sin metodología.
- Un modelo es un algoritmo entrenado.
- Depende de los patrones en los datos
- No hay toma de requisitos para una salida esperada.
- Incertidumbre
- Solo se tiene un control indirecto.

Agenda

- Introducción
- Comparativo de Metodologías
- Aproximación Metodológica
- Conclusiones

Comparativo de Metodologías



- **Hitos asociados al Desarrollo de Software**
- **Metodologías de Minería de Datos o Analítica**
- **Otros hitos de relevancia dentro del desarrollo DataOps**

Área Curricular de Sistemas e Informática
Facultad de Minas
Sede Medellín
2020

CRISP-DM - 1997

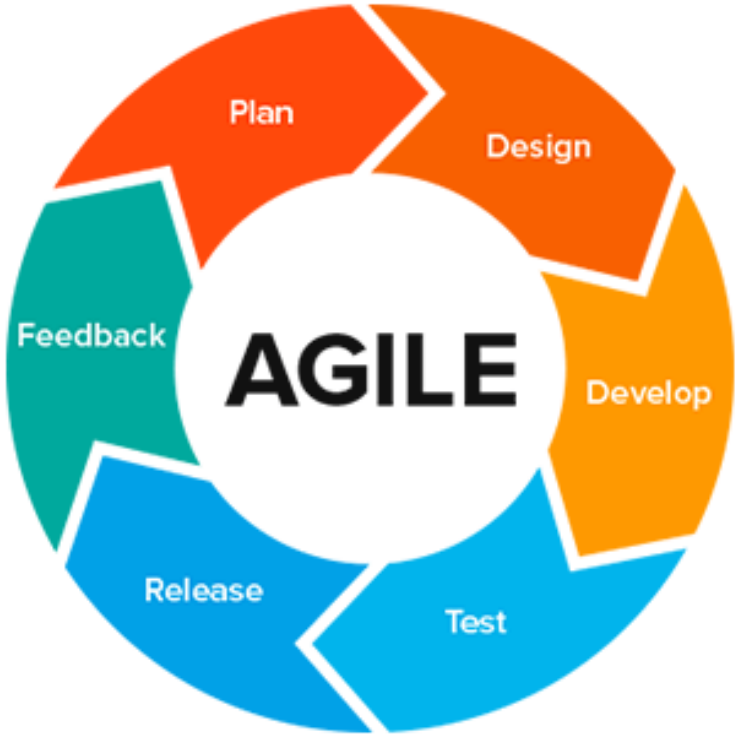
1997
IBM

- Estándar de Industria
- El ciclo de vida de un proyecto de minería de datos consta de seis fases.
- La secuencia de las fases no es lineal. Siempre se requiere avanzar y retroceder entre las diferentes fases.
- Depende del resultado obtenido en cada fase, se decide con cual fase se debe continuar.

UNIVERSIDAD NACIONAL DE COLOMBIA

Metodologías Ágiles - 2001

2001
Beedle



- Mecanismo iterativo para producir software.
- Ciclo de diseño y prueba de código al menos una vez al día
- Concentración de equipos interdisciplinarios.
- Garantía de calidad y aumento de frecuencia de entrega.
- La entrega de software está cubierto por DevOps.

Metodologías Ágiles - 2001

A Favor:

- Aceptan los cambios: los modelos cambian con los datos.
- Manejan la incertidumbre al dividir en pequeños pasos.
- Tiene buenas prácticas para la gestión del desarrollo de SW.

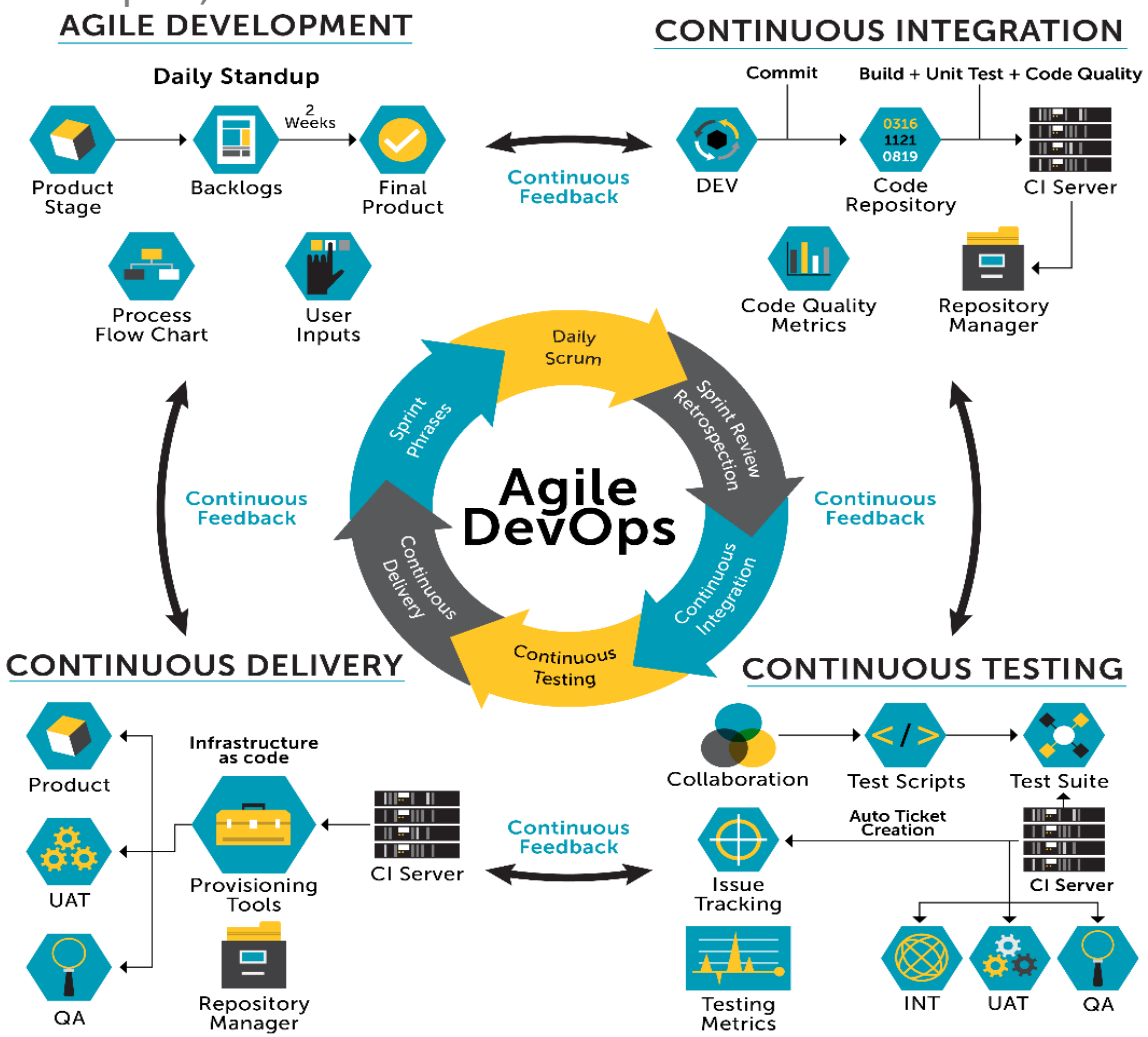
En Contra:

- Las ceremonias y métricas no se ajustan al tipo de proyecto.
- Los compromisos con los pequeños pasos son difíciles de hacer.
- Su forma de planeación no funciona en ML.

DevOps - 2009

2009

Allspaw, Hammond



- Proceso para el desarrollo de software que hace énfasis en la colaboración, comunicación, ***automatización*** e integración

- Reduce ***tiempos*** entre la construcción y la implementación en un ambiente productivo, asegurando la ***calidad***

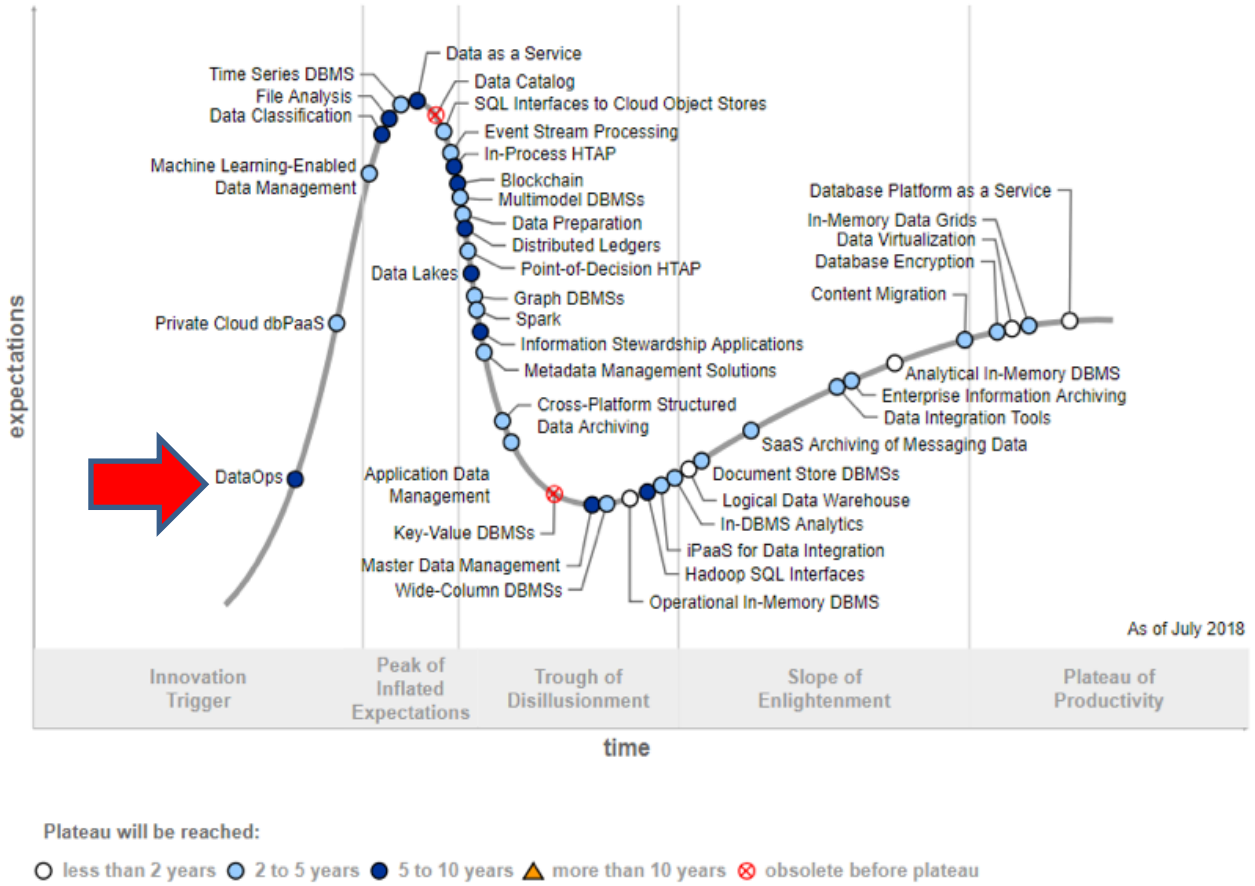
- Uso de prácticas continuas para la ***creación*** y destrucción de ***entornos*** a demanda.

UNIVERSIDAD NACIONAL DE COLOMBIA

DataOps – 2015

2015

Palmer



- ***DataOps*** es una práctica de administración de datos en colaboración centrada en mejorar la comunicación, **integración** y **automatización** de los flujos de datos entre los administradores de datos y los consumidores en toda la organización.

- Al igual que DevOps, ***DataOps*** no es un dogma rígido, sino una **práctica basada en principios** que influye en cómo se pueden proporcionar y actualizar los datos para satisfacer las necesidades de los consumidores de datos de la organización.

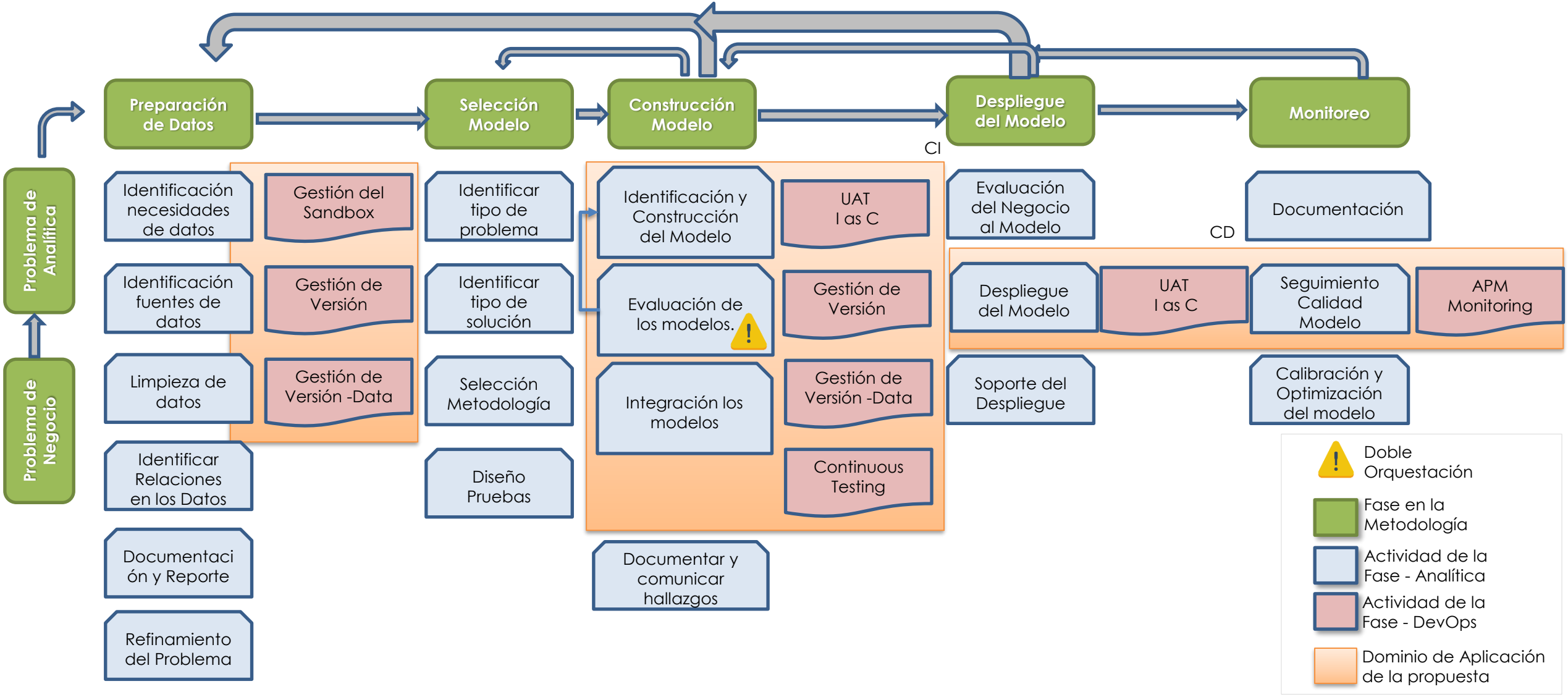
- Actualmente, las soluciones disponibles son propuestas de industria basadas en **software comercial** y casi siempre licenciado.

UNIVERSIDAD NACIONAL DE COLOMBIA

Agenda

- Objetivo de un proyecto de ML
- Desafíos en los proyectos de ML
- Comparativo de Metodologías
- Aproximación Metodológica
- Conclusiones

Aproximación Metodológica



Aproximación Metodológica

Doble Orquestación

Prueba Integral

Código

pytest

sonarqube

Prueba unitaria
Prueba estática

Acceso

SQL/ETL

Modelo

Visualización

Reporte

Prueba Funcional
Prueba Rendimiento
Prueba de Humo

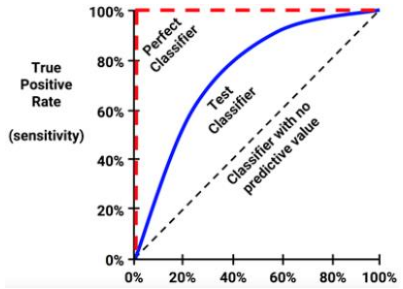
Ambiente Productivo

Datos / Modelo

Cantidad de Registros y Campos

$\Sigma A \geq \Sigma B$

Comparación de Datos



AIC

BIC

$MAPE$

Evaluación de indicadores de Analítica

$\sigma \quad u \quad P(0)$

Control Estadístico

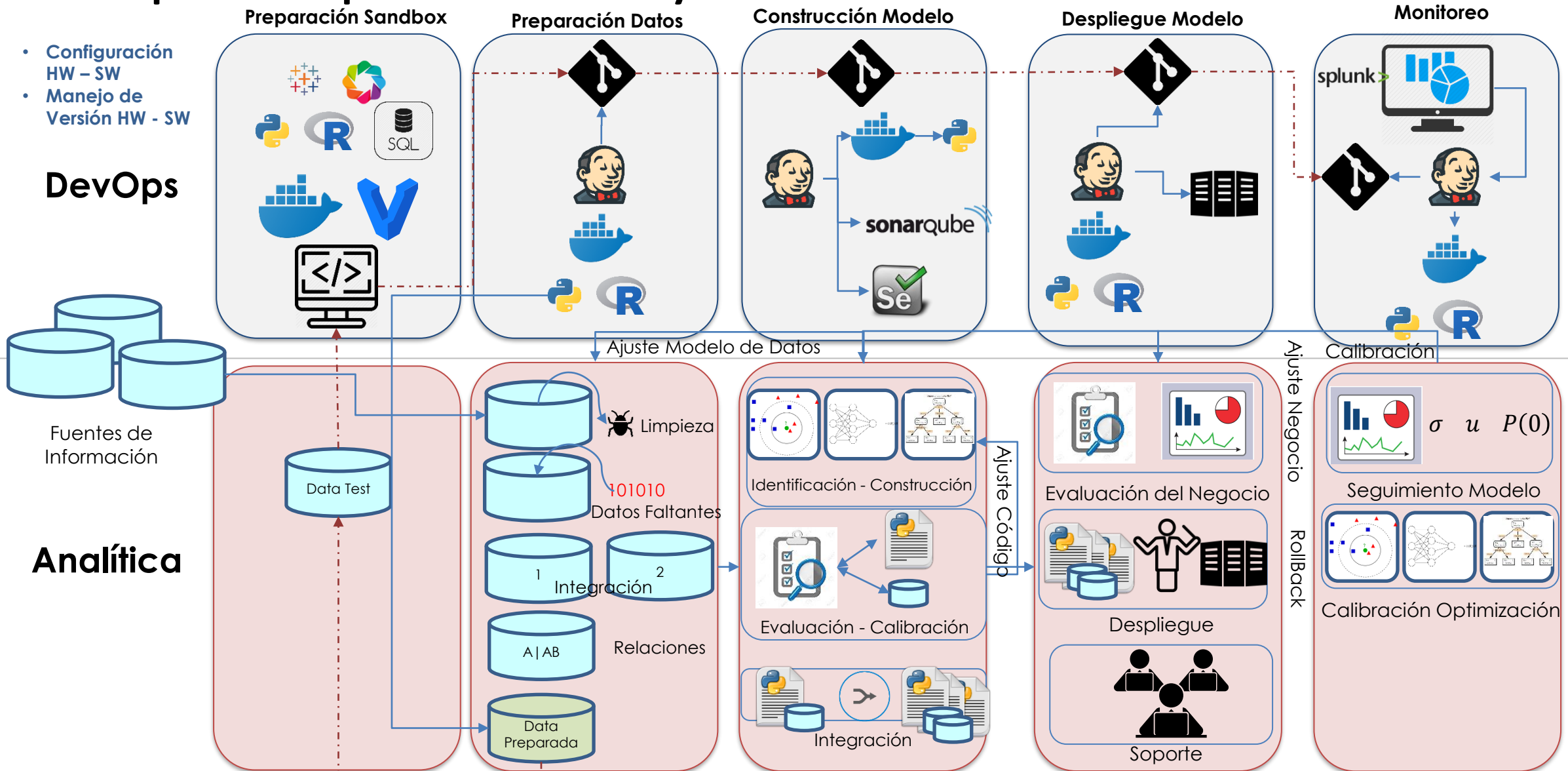
Aproximación Metodológica

Pipeline para un Proyecto de ML

- Configuración HW – SW
- Manejo de Versión HW - SW

DevOps

Analítica



Agenda

- Introducción
- Comparativo de Metodologías
- Aproximación Metodológica
- Conclusiones

Conclusiones

- Para la preparación de datos, construcción y despliegue, es necesario la definición de las herramientas que hagan resuo de artefactos de infraestructura en pro de automatizar y liberar tiempo.
- La realización de la ***doble orquestación***, constituye la principal diferencia de la aproximación metodológica respecto a un proyecto de desarrollo tradicional de software con DevOps.
- Es fundamental el uso de contendores, que permitan manejar de manera centralizada la configuración de entornos, reducir tiempos de despliegue y eliminar reprocesos por mantenimiento.
- Una buena definición de herramientas, scripts, archivos de configuración de los contenedores usados, y la definición de la evaluación en la doble orquestación es una inversión segura para que se debe ejecuciones del pipeline limpias y rápidas.

Sección, Facultad, Oficina, Departamento…

Dirección:

Carrera 65 Nro. 59A – 110 Bloque xx – Oficina xxx
Medellín, Colombia
(+57 4) 430 90 00 ext. xxxxxxxx
xxxxxxxxxxxxxxxxxxxx@unal.edu.co