

Buenas prácticas y estándares en Data Mining y en Ciencia y Tecnología

Lista de deseos:

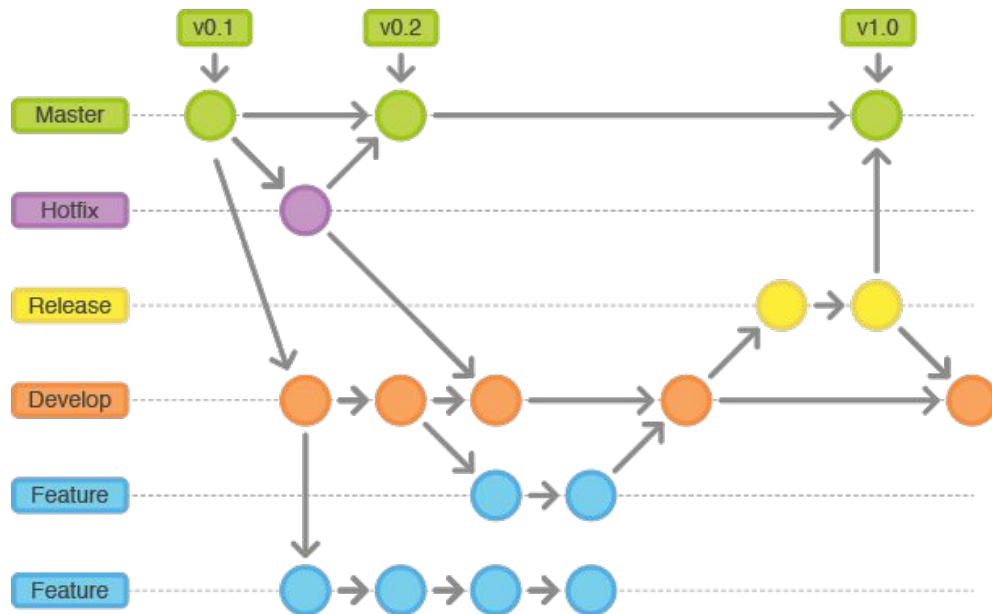
- Objetivos o hipótesis a priori
 - Poder empezar un proyecto sabiendo que se quiere
- Transparencia de procedimientos
 - Entender cada paso para poder documentar los más claro los procedimientos
- Reproducibilidad de resultados
 - No dejar nada librado al azar (en la medida que se pueda)
- Integridad y persistencia de los datos
 - No adulterar los datos de forma irreversible
- Automatización de tareas
 - Evitar si es posible las tareas que no pueden ejecutarse de forma automática
- Reporte de resultados adecuados
 - Visualización, legibilidad

Herramientas:

- Git
 - Control de versiones
- Latex
 - Sistema para crear documentos
- CRISP-DM:
 - Cross Industry Standard Process for Data Mining
- Predictive Model Markup Language (PMML)
 - Formato estandar para intercambio de modelos (viejo)
- Portable Format for Analytics (PFA)
 - Formato estandar para intercambio de modelos (nuevo)
- Open Neural Network Exchange (ONNX)
 - Formato estandar para intercambio de modelos (más nuevo)

Git

- El sistema de control de versiones más usado
- Creado por Linus Torvalds, creador de Linux
- Descentralizado
- Flujos de trabajo no lineales
- Muchos hosts gratuitos
 - github.com
 - gitlab.com
 - bitbucket.com



Git: uso típico

- Crear un repositorio remoto
- *Clonar* el repositorio en la computadora
- *Agregar* archivos para que sean seguidos por el repositorio
- Realizarle cambios al archivo
- *Comitear* los cambios indicando que se hizo
- *Pushear* los archivos al repositorio remoto
- Crear una *rama* nueva si se quiere probar algo nuevo y no romper todo
- Si la prueba funcionó incluir los cambios en la *rama master*

Github Desktop: repositorio nuevo

Create a new repository

×

Name

repository name

Description

Local path

/home/miles/Documents/GitHub

Choose...

☐ Initialize this repository with a README

Git ignore

None ▼

License

None ▼

Create repository

Cancel

Github Desktop: publicar

- Mismo nombre al repositorio en github que en la computadora

No local changes

You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.



Publish your repository to GitHub

This repository is currently only available on your local machine. By publishing it on GitHub you can share it, and collaborate with others.

Always available in the toolbar for local repositories or **Ctrl** **P**

Publish repository

Publish repository



GitHub.com

GitHub Enterprise Server

Name

dmcyt

Description

☒ Keep this code private

Organization

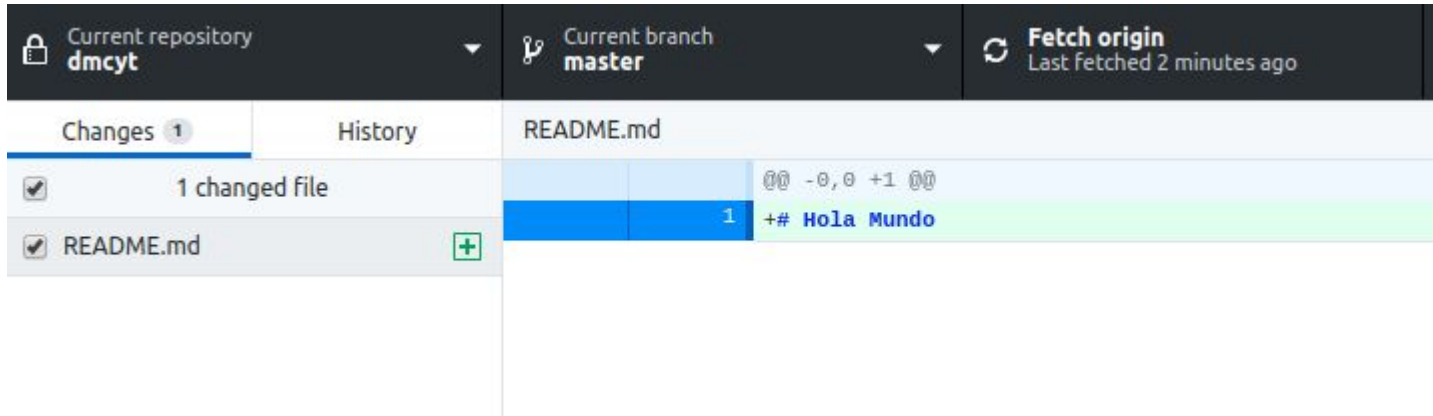
None

Publish repository

Cancel

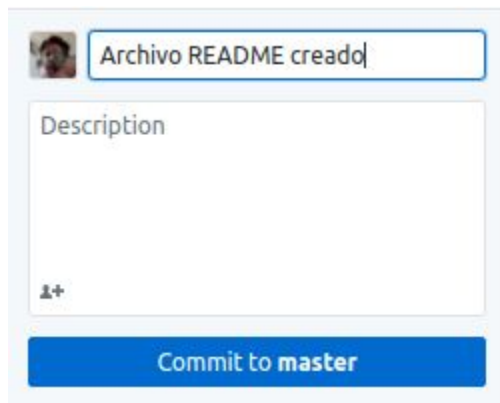
Github Desktop: agregar contenido

- Archivo README.md



Github Desktop: commitear

- Incluir descripción de los cambios hechos



A screenshot of the GitHub Desktop commit dialog box. At the top left is a small circular profile picture. To its right is a text input field containing the text "Archivo README creado". Below this is a large text area with the placeholder text "Description". At the bottom left of the text area is a small icon of a person with a plus sign. At the bottom of the dialog is a blue button with the text "Commit to master".

Github Desktop: pushear

- Enviar al repositorio remoto los commits

No local changes

You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.



Push 1 commit to the origin remote

You have one local commit waiting to be pushed to GitHub

Always available in the toolbar when there are local commits waiting to be pushed

or **Ctrl** **P**

Push origin

LaTeX:

```
1 \documentclass{article}
2 \usepackage[utf8]{inputenc}
3
4 \title{DMCyT}
5 \author{pablo.riera }
6 \date{August 2019}
7
8 \usepackage{natbib}
9 \usepackage{graphicx}
10
11 - \begin{document}
12
13 \maketitle
14
15 - \section{Introduction}
16 There is a theory which states that if ever anyone discovers exactly what the Universe is for and
17 why it is here, it will instantly disappear and be replaced by something even more bizarre and
18 inexplicable.
19 There is another theory which states that this has already happened.
20
21 - \begin{figure}[h!]
22 \centering
23 \includegraphics[scale=1.7]{universe}
24 \caption{The Universe}
25 \label{fig:universe}
26 \end{figure}
27
28 - \section{Conclusion}
29 ``I always thought something was fundamentally wrong with the universe''
30 \citep{adams1995hitchhiker}
31
32 \bibliographystyle{plain}
33 \bibliography{references}
34 \end{document}
```

DMCyT

pablo.riera

August 2019

1 Introduction

There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another theory which states that this has already happened.



Figure 1: The Universe

2 Conclusion

“I always thought something was fundamentally wrong with the universe” [1]

References


[1] D. Adams. *The Hitchhiker's Guide to the Galaxy*. San Val, 1995.


Overleaf: sacar una cuenta


Log in to Overleaf


Log in with your email

or

 Log in with IEEE

 Log in with Google

 Log in with Twitter

 Log in with ORCID

First time here as a ShareLaTeX user?

You can now log in to your ShareLaTeX account through Overleaf. [Find out more.](#)

Don't have an account? [Register](#)

[Forgot your password?](#)

LaTeX: básicos

- Tipo de artículo
 - `\documentclass{article}`
- Título, autor, fecha, etc
 - `\title{DMCyT}`
`\author{pablo.riera }`
`\date{August 2019}`
- Documento
 - `\begin{document}`
`\maketitle`
`\end{document}`
- Secciones
 - `\section{Introducción}`
 - `\subsection{Temas}`
- Bibliografía
 - `\bibliographystyle{plain}`
 - `\bibliography{references}`

CRISP-DM: CRoss Industry Standard Process for Data Mining

- Es un modelo de proceso de data-mining que es independiente de la herramienta, la aplicación y la industria.
- La versión 1.0 de la guía se publicó en 2000. El consorcio que promueve el uso de CRIPS actualmente está inactivo.
- En 2015, IBM lanzó una nueva metodología llamada Método Unificado de Analytics Solutions para Minería de Datos / Análisis Predictivo (también conocido como ASUM-DM) que refina y extiende CRISP-DM.

Metodología Jerárquica (modularizar, divide y venceras)

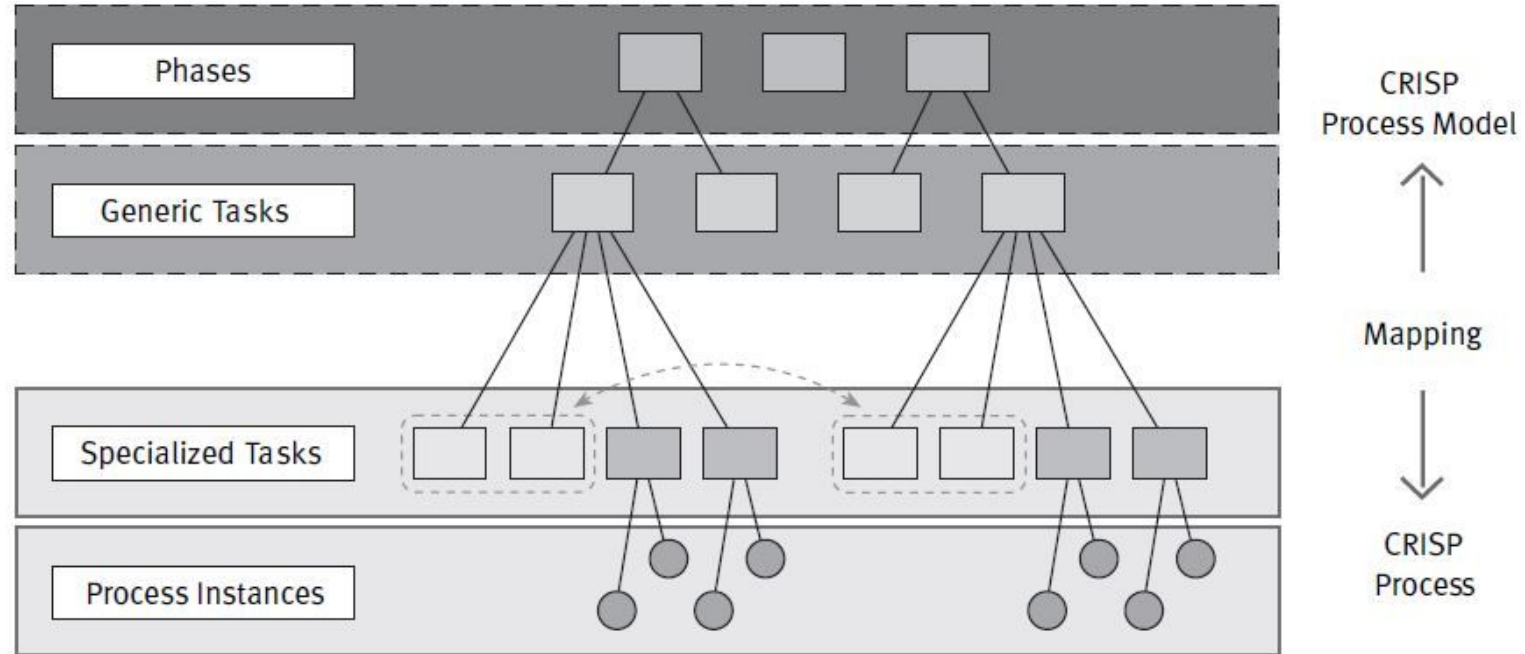


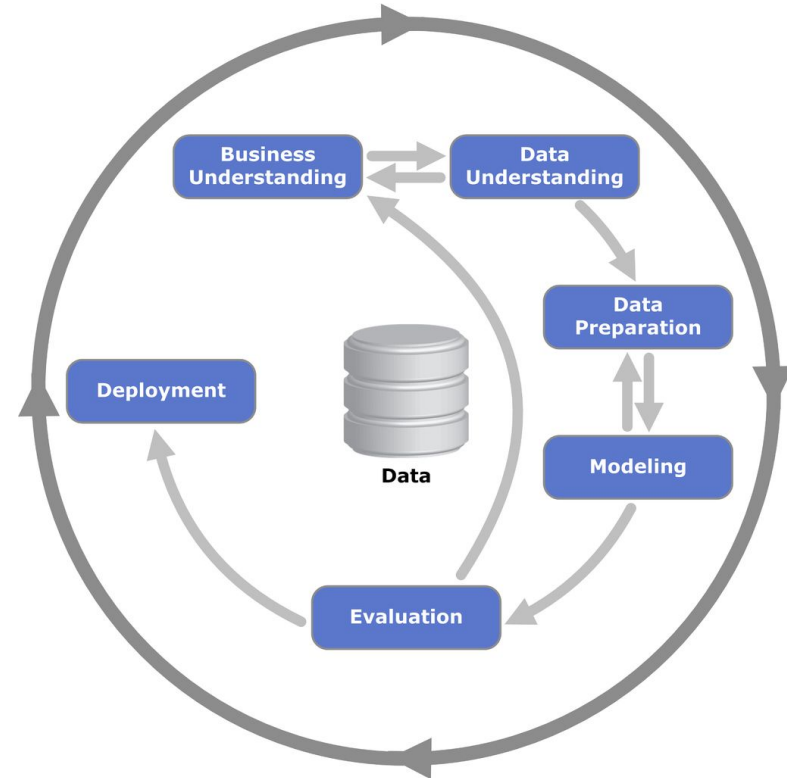
Figure 1: Four level breakdown of the CRISP-DM methodology

Contextos de Data Mining:

- Dominio de aplicación:
 - Area específica de aplicación del proyecto de data mining (p.ej. bioinformática)
- Tipo de problema:
 - Objetivos del proyecto de data-mining (p.ej., clasificación)
- Aspectos técnicos:
 - Temas específicos de data-mining que se refieren a las dificultades y particularidades del proyecto (p.ej., tipos de datos en una base de datos)
- Técnicas y herramientas:
 - Herramientas y técnicas de data mining que se utilizan en el proyecto (p.ej, k-medias, PAM)

CRISP-DM: Modelo de Referencia

- Contiene las fases del proyecto, sus tareas respectivas y algunas relaciones entre tareas. A este nivel no es posible identificar todas las relaciones
 1. Comprensión del dominio
 2. Comprensión de los datos
 3. Preparación de los datos
 4. Modelado
 5. Evaluación
 6. Despliegue / implementación

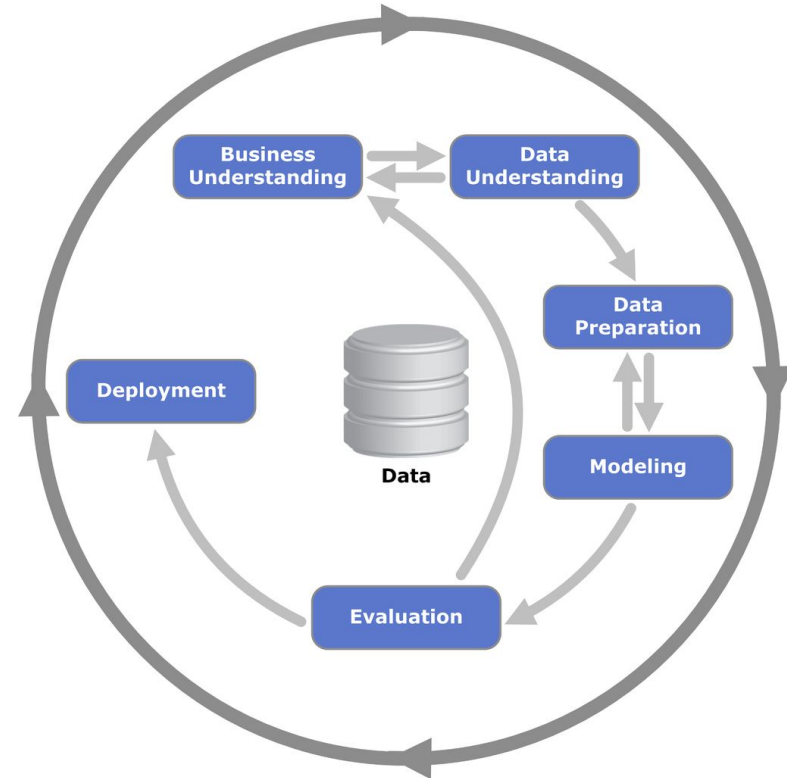


CRISP-DM: Comprensión del dominio

1. Determinar objetivos
 - 1.1. Información general del dominio
 - 1.2. Definir objetivos
 - 1.3. Definir el criterio de éxito
2. Evaluar la situación
 - 2.1. Recursos
 - 2.2. Requerimientos, supuestos, condicionantes
 - 2.3. Condiciones de riesgo y contingencias
 - 2.4. Terminología
 - 2.5. Determinar costos y beneficios
3. Objetivos de data mining
 - 3.1. Determinar los objetivos
 - 3.2. Definir el criterio de éxito
4. Producir el plan del proyecto
 - 4.1. Redacción del proyecto
 - 4.2. Evaluación inicial de técnicas y herramienta

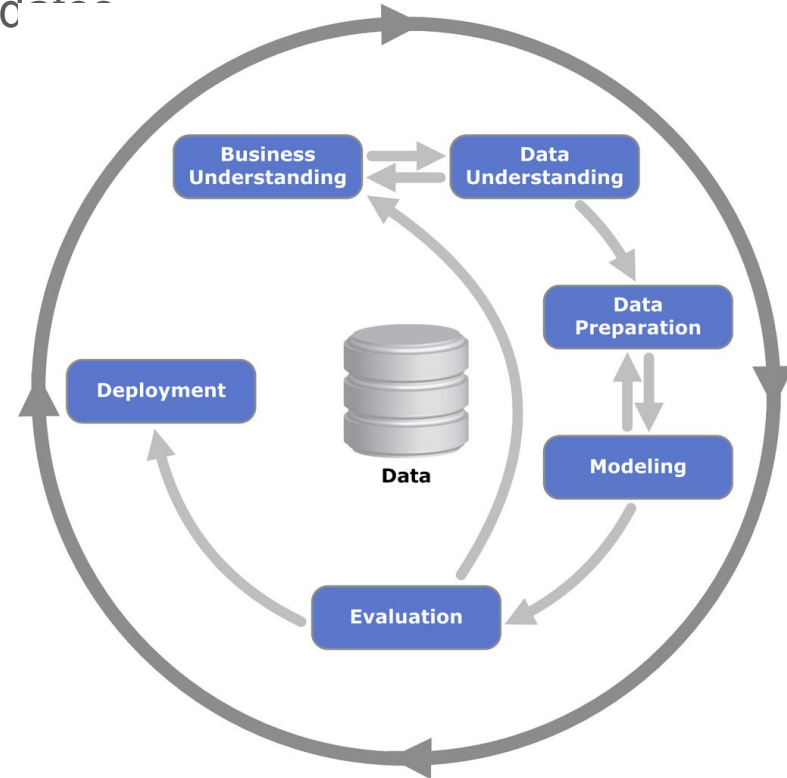
CRISP-DM: Comprensión de los datos

1. Colección inicial de datos
 - Informe inicial de colección de datos
2. Describir los datos
 - Informe de descripción de datos
3. Exploración de datos
 - Informe de exploración de datos
4. Verificar la calidad de los datos
 - Informe de calidad de los datos



CRISP-DM: Preparación de los datos

1. Obtener / Seleccionar el conjunto inicial de datos
2. Limpiar datos
3. Construir datos
 - Crear atributos derivados
 - Crear nuevos registros
 - Aplicar transformaciones
4. Integración de los datos
5. Formateo de los datos



CRISP-DM: Modelado

1. Seleccionar la técnica de modelado
2. Generar el diseño de prueba
 - Crear conjuntos de entrenamiento y de prueba
3. Construir el modelo
 - Determinar parámetros del modelo
 - Modelar
 - Describir el modelo
4. Analizar el modelo
 - Evaluación (comportamiento, ranking de modelos)
 - Reajuste de los parámetros del modelo



CRISP-DM: Evaluación

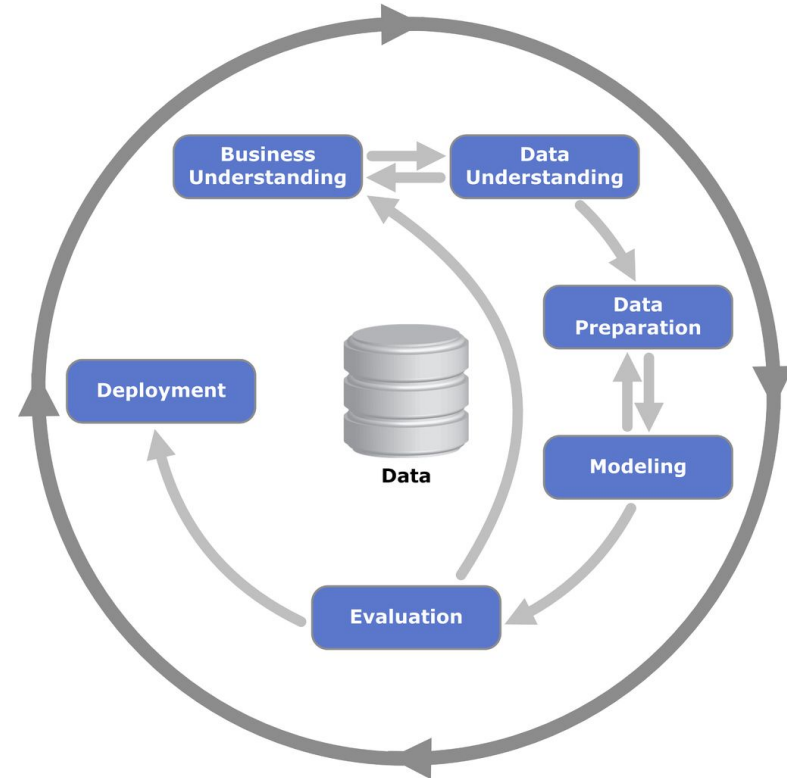
1. Evaluación de resultados

- Análisis de los resultados de DM
- Selección de modelos

2. Proceso de revisión

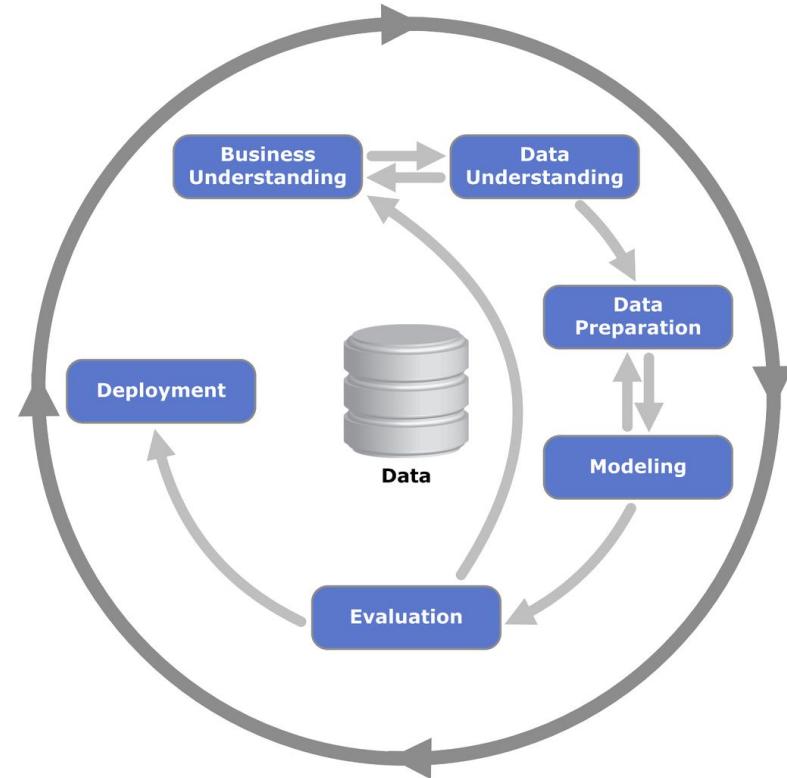
3. ¿Próximos pasos?

- Lista de posibles acciones
- Decisiones



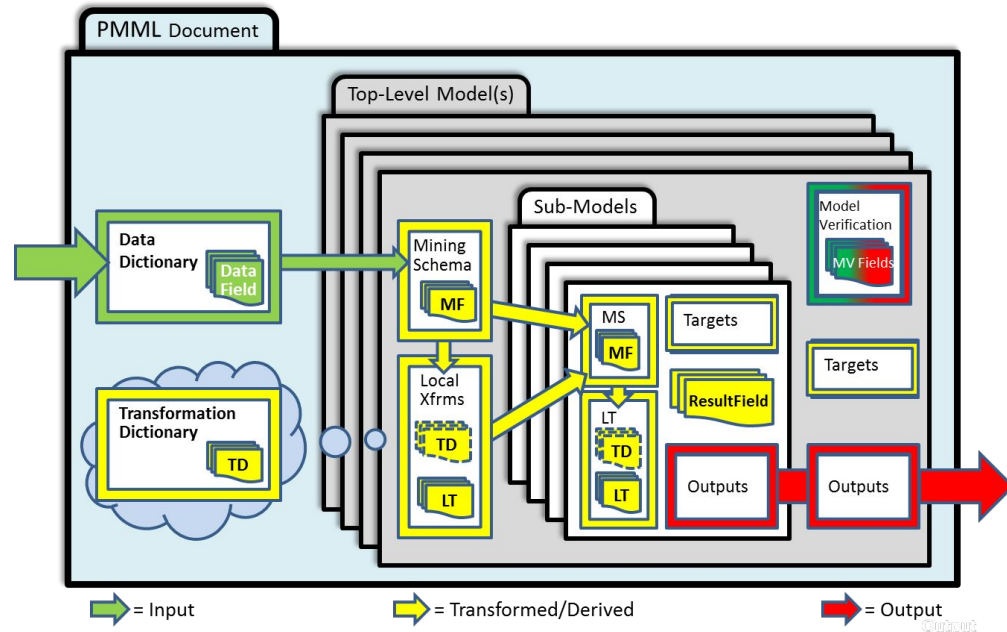
CRISP-DM: Despliegue / Implementación

1. Plan de despliegue / implementación
 - Análisis de los resultados de DM
 - Selección de modelos
2. Plan de monitoreo y mantenimiento
 - Informe de descripción de datos
3. Preparación del informe final
4. Revisión del proyecto



Predictive Model Markup Language (PMML)

- Lenguaje de marcado de texto XML desarrollado por el Data Mining Group (DMG) para proveer a las aplicaciones una manera de definir modelos relacionados con los análisis predictivos y la minería de datos para compartir estos modelos entre las aplicaciones PMML.



Predictive Model Markup Language (PMML)

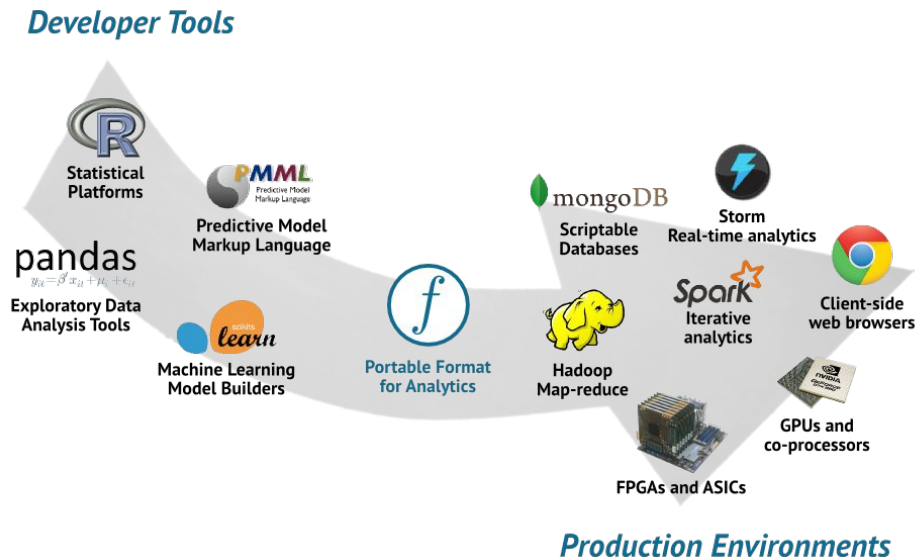
- **Header:** información general, información de copyright del modelo, descripción e información sobre la aplicación utilizada para generar el modelo, el nombre y la versión.
- **Data Dictionary:** tipos de variables (contínuos, categóricos, ordinales), rangos, válidos, inválidos y faltantes
- **Data Transformation:** normalización, discretización, asignación, agregación
- **Model:** definición, nombre, atributos
- **Mining Scheme:** datos usados para modelar, valores predichos
- **Target:** modificaciones post-procesado, escalado
- **Output:** nombres de las salidas

Ejemplo (PMML)

```
<?xml version="1.0" encoding="UTF-8"?>
<PMML xmlns="http://www.dmg.org/PMML-3_2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="3.2" xsi:schemaLocation="http://www.dmg.org/PMML-3_2 http://www.dmg.org/v3-2/pmml-3-2.xsd">
  <Header copyright="Copyright (c) 2012 DMG" description="KMeans cluster model">
    <Extension name="user" value="DMG" extender="Rattle/PMML" />
    <Application name="Rattle/PMML" version="1.2.29" />
    <Timestamp>2012-09-27 13:19:09</Timestamp>
  </Header>
  <DataDictionary numberOfFields="4">
    <DataField name="sepal_length" optype="continuous" dataType="double" />
    <DataField name="sepal_width" optype="continuous" dataType="double" />
    <DataField name="petal_length" optype="continuous" dataType="double" />
    <DataField name="petal_width" optype="continuous" dataType="double" />
  </DataDictionary>
  <ClusteringModel modelName="KMeans_Model" functionName="clustering" algorithmName="KMeans: Hartigan and Wong" modelClass="centerBased" numberOfClusters="3">
    <MiningSchema>
      <MiningField name="sepal_length" usageType="active" />
      <MiningField name="sepal_width" usageType="active" />
      <MiningField name="petal_length" usageType="active" />
      <MiningField name="petal_width" usageType="active" />
    </MiningSchema>
    <ComparisonMeasure kind="distance">
      <squaredEuclidean />
    </ComparisonMeasure>
    <ClusteringField field="sepal_length" compareFunction="absDiff" />
    <ClusteringField field="sepal_width" compareFunction="absDiff" />
    <ClusteringField field="petal_length" compareFunction="absDiff" />
    <ClusteringField field="petal_width" compareFunction="absDiff" />
    <Cluster name="1" size="24">
      <Array n="4" type="real">6.88333333333333 3.09166666666667 5.8375 2.12916666666667</Array>
    </Cluster>
    <Cluster name="2" size="33">
      <Array n="4" type="real">5.06060606060606 3.47272727272727 1.45454545454545 0.25454545454545</Array>
    </Cluster>
    <Cluster name="3" size="48">
      <Array n="4" type="real">5.93125 2.75416666666667 4.46041666666667 1.45416666666667</Array>
    </Cluster>
  </ClusteringModel>
</PMML>
```

Portable Format for Analytics (PFA)

- El Portable Format for Analytics (PFA) es un formato de intercambio de modelos predictivo basado en JSON desarrollado por el Data Mining Group
- PFA proporciona una forma para que las aplicaciones analíticas describan e intercambien modelos predictivos producidos por algoritmos de aprendizaje automático.



<https://www.kdnuggets.com/2016/01/portable-format-analytics-models-production.html>

Portable Format for Analytics (PFA)

- Ejemplo:

```
{ "input": "double",  
  "output": "double",  
  "action": [  
    { "+": ["input", 100] }  
  ]  
}
```

1	1	input (JSONS)
2	2	
3	3	

1	input: double	PFA document (YAML)
2	output: double	
3	action:	
4	- {+: [input, 100]}	

1	101.0	output (TEXT)
2	102.0	
3	103.0	