La **Ciencia de** *los* **Datos**

Mg. Ing. Karina B. Eckert

Agenda

1. Objetivo
2. Introducción
3. Antecedentes y disciplinas vinculadas
4. Ciencia de Datos
    1. Metodologías
    2. Datos y almacenamiento
    3. Perfiles/Roles
    4. Lenguajes y Herramientas
5. Aplicaciones reales

OBJETIVO

Objetivo

Dar a conocer los conceptos centrales vinculados a la ciencia de datos: principios, metodologías, técnicas, algoritmos, roles, herramientas de ciencia de datos…

INTRODUCCIÓN

Motivación

Introducción

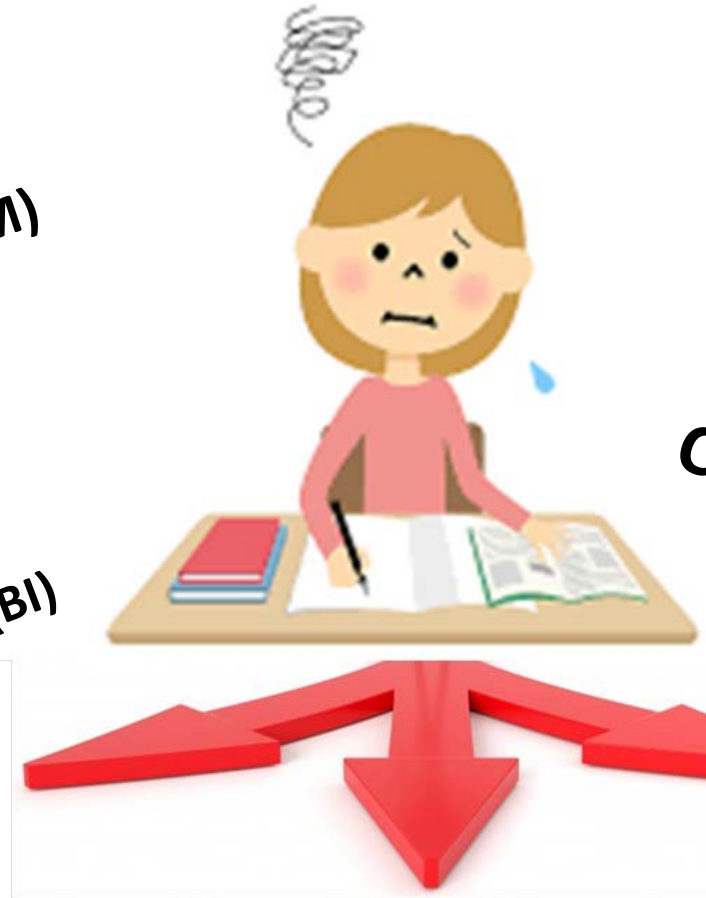
Información adecuada, en el lugar y momento oportuno

Incrementa la Efectividad

"Torturar los datos hasta que ellos confiesen"

CD/MD

Introducción



Minería de Datos (DM)

Análisis de Datos

Ciencia de Datos (DS)

Inteligencia de Negocio (BI)

?

?

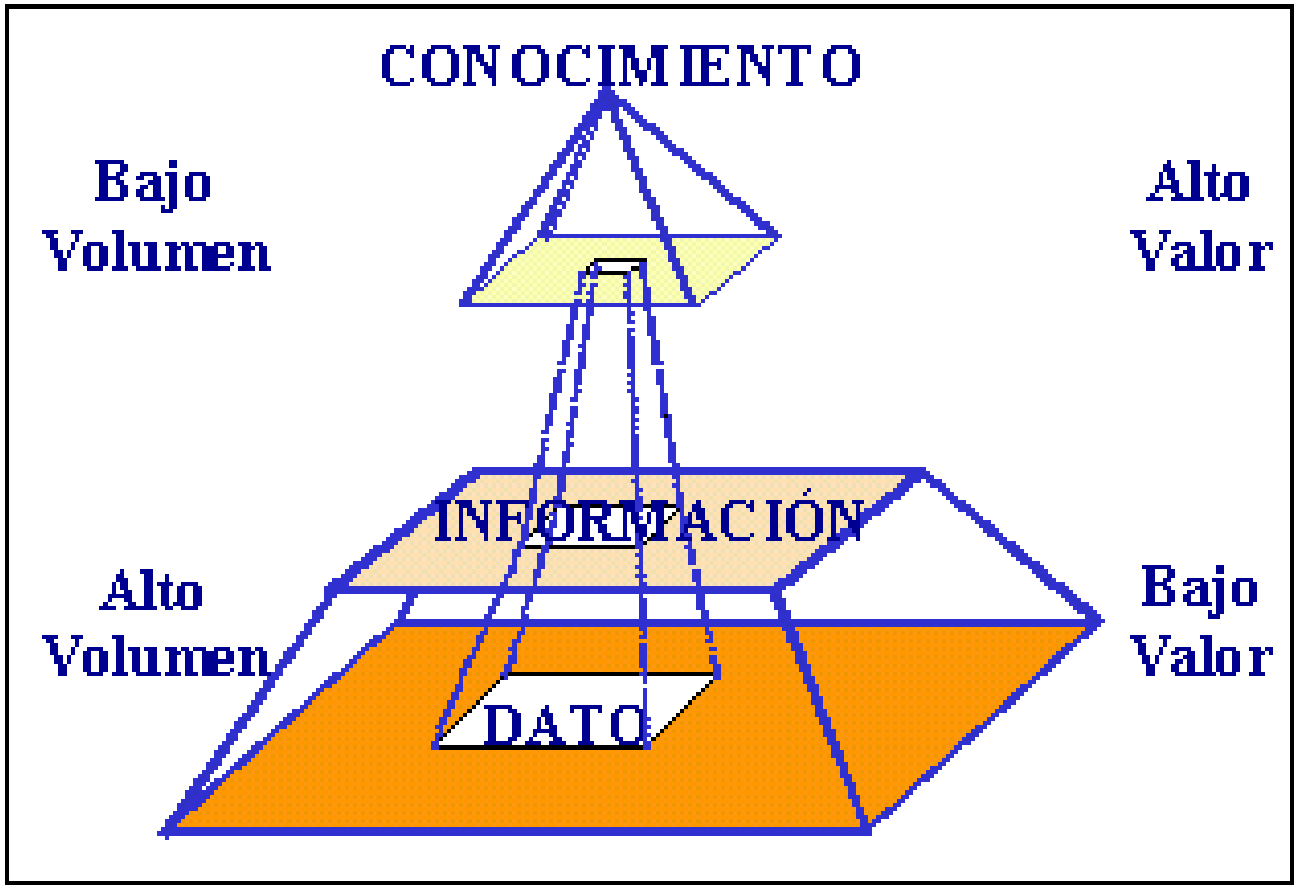
?

Big Data

Extracción de Conocimiento en Base de Datos (KDD)

Introducción

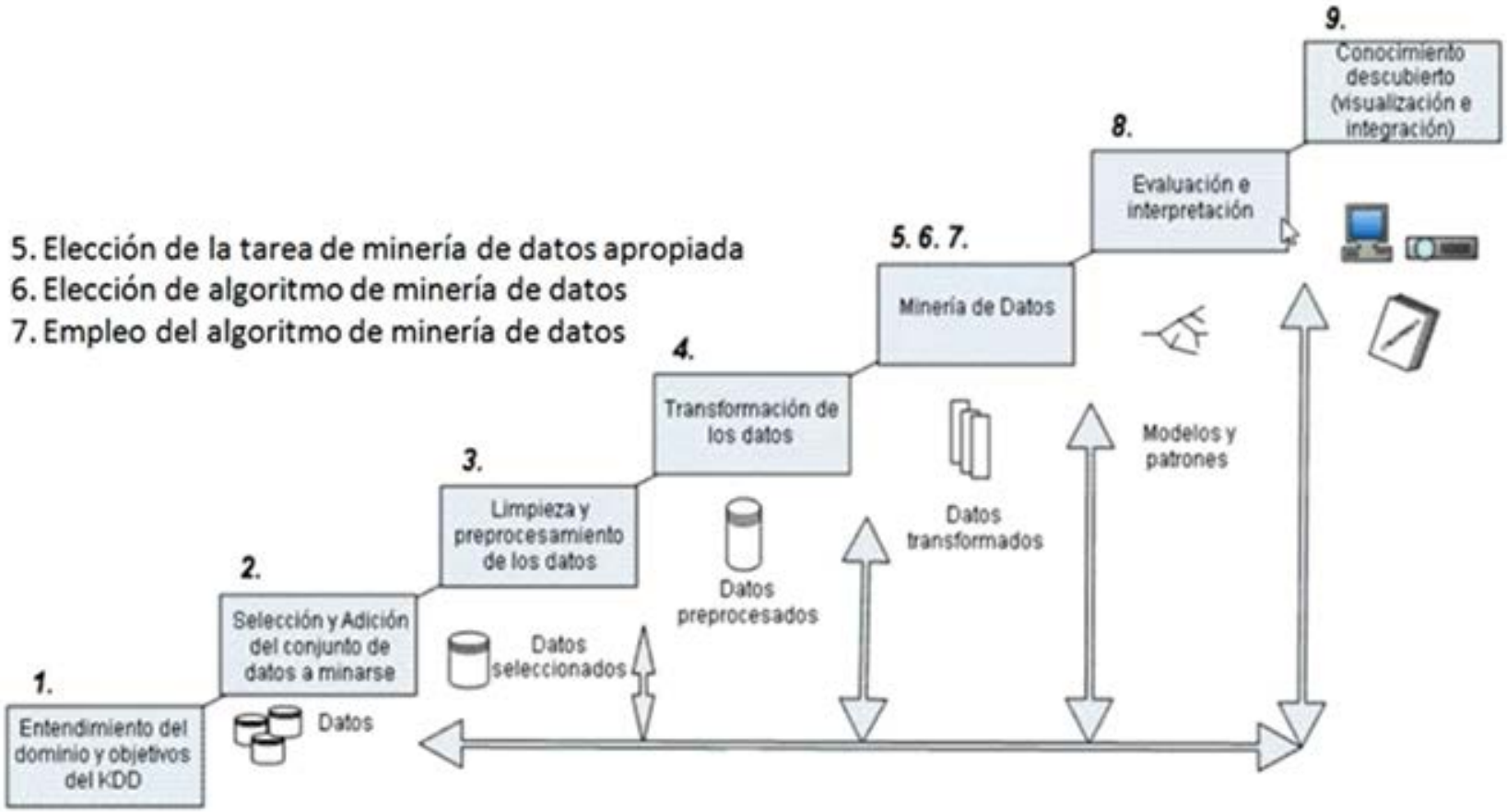
Introducción



Pirámide del Conocimiento

Molina, 1998

Introducción



5. Elección de la tarea de minería de datos apropiada
6. Elección de algoritmo de minería de datos
7. Empleo del algoritmo de minería de datos

KDD: Knowledge Discovery in Databases
KDD: Knowledge Discovery and DataMining

Introducción

"Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos"

Metas:

- Procesar automáticamente grandes cantidades de datos.
- Identificar patrones significativos y relevantes.
- Presentar los modelos como conocimiento innovador y apropiado para satisfacer las metas del usuario.

Fayyad y otros. 1996

Introducción

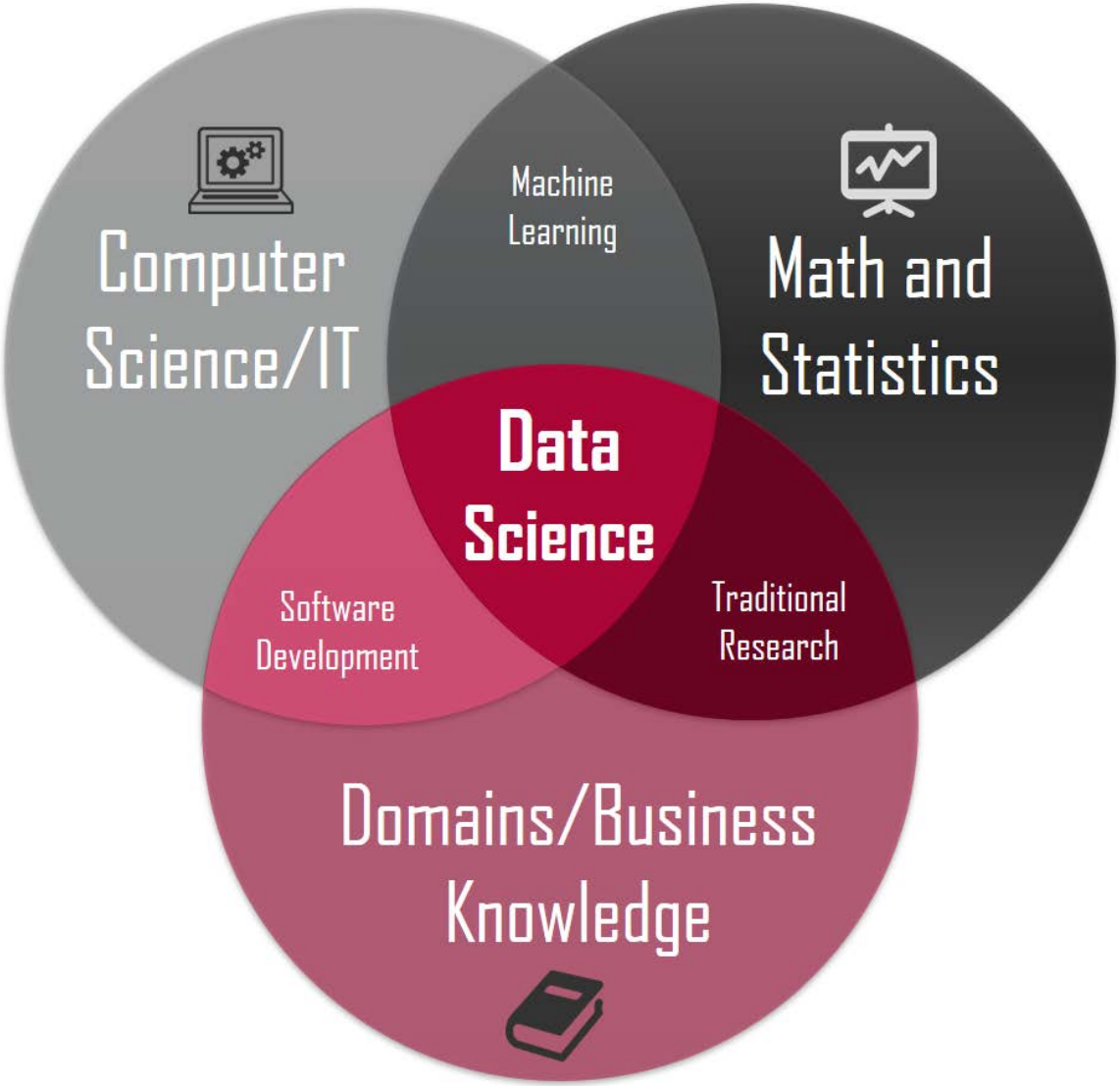
Inteligencia de Negocio

- Es el conjunto de procesos, metodologías, estrategias, aplicaciones y tecnologías que facilitan la obtención rápida y sencilla de todos los datos generados por una empresa para su análisis e interpretación, de manera que puedan ser aprovechados para la toma de decisiones y se conviertan en conocimiento para los responsables del negocio.

- BI, es analizar los datos históricos para comprender lo que pasó. En algunos casos permite entender una tendencia.

Introducción

DS es un conjunto de principios fundamentales que apoyan y guían la extracción de información y conocimiento a partir de los datos; incluye diversas metodologías, técnicas, algoritmos y herramientas que facilitan el procesamiento avanzado y automático de los mismos; permitiendo identificar información relevante y estratégica, que a simple vista no es detectada.



Introducción

"El Big Data es el análisis masivo de datos. Una cuantía de datos, tan sumamente grande, que las aplicaciones de software de procesamiento de datos que tradicionalmente se venían usando no son capaces de capturar, tratar y poner en valor en un tiempo razonable."

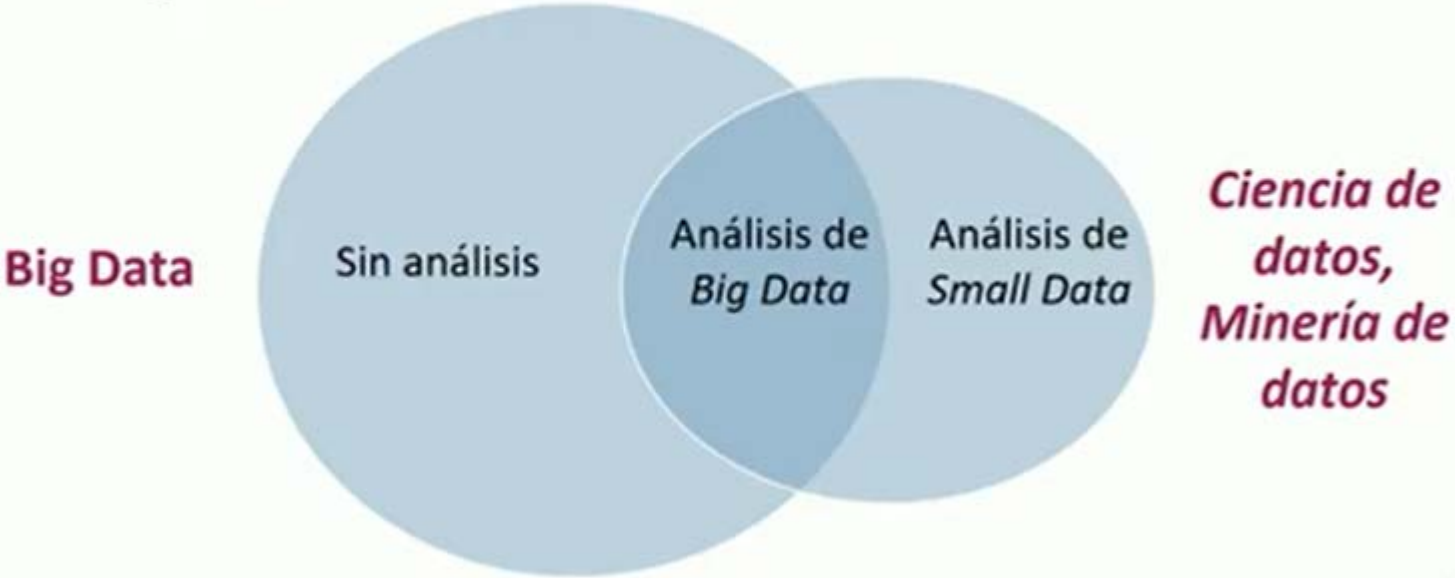
Herramientas

Apache Hadoop	MongoDB	Cassandra
Elasticsearch	Apache Spark	Lenguaje R
Apache Storm	Python	Apache Drill
	Apache	Oozie

https://www.masterbigdataucm.com/que-es-big-data/1

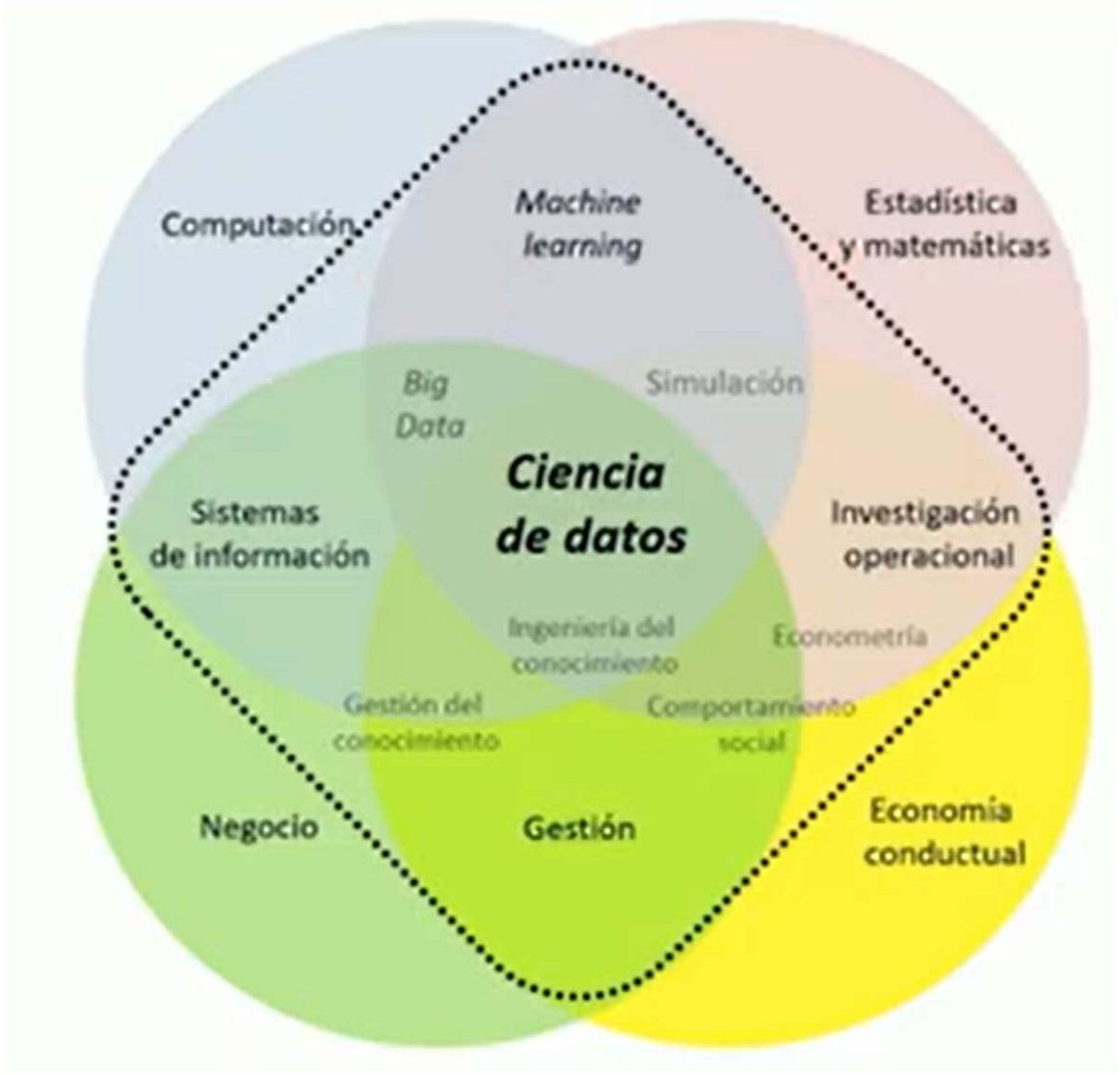
Introducción

- No toda la ciencia de datos es Big Data.
- No todo el Big Data es ciencia de datos.



Big Data y Ciencia de Datos

Ciencia de Datos



CIENCIA DE DATOS

Ciencia de Datos

- Campo inter/trans disciplinar que involucra
    - Métodos científicos
    - Procesos
    - Sistemas

  Para extraer conocimiento o

  mejorar el entendimiento de los datos (estructurados o no)

- Involucra diversas disciplinas para los diferentes tipos de análisis, como:
  - Minería de datos
  - Aprendizaje Automático
  - Estadística computacional

Ciencia de Datos

- Data Science es la **resolución** a los **problemas de negocios/organizaciones** a través de las **matemáticas**, la **programación** y el **método científico** que implica la creación de **hipótesis** , **experimentos** y **pruebas** a través del **análisis de datos** y la generación de **modelos predictivos**.
- Es responsable de transformar estos problemas en **preguntas bien planteadas** que también puedan responder a la hipótesis inicial de una manera **creativa**. También debe incluir la **comunicación efectiva** de los resultados obtenidos y cómo la solución **agrega valor** a la Empresa/Organización

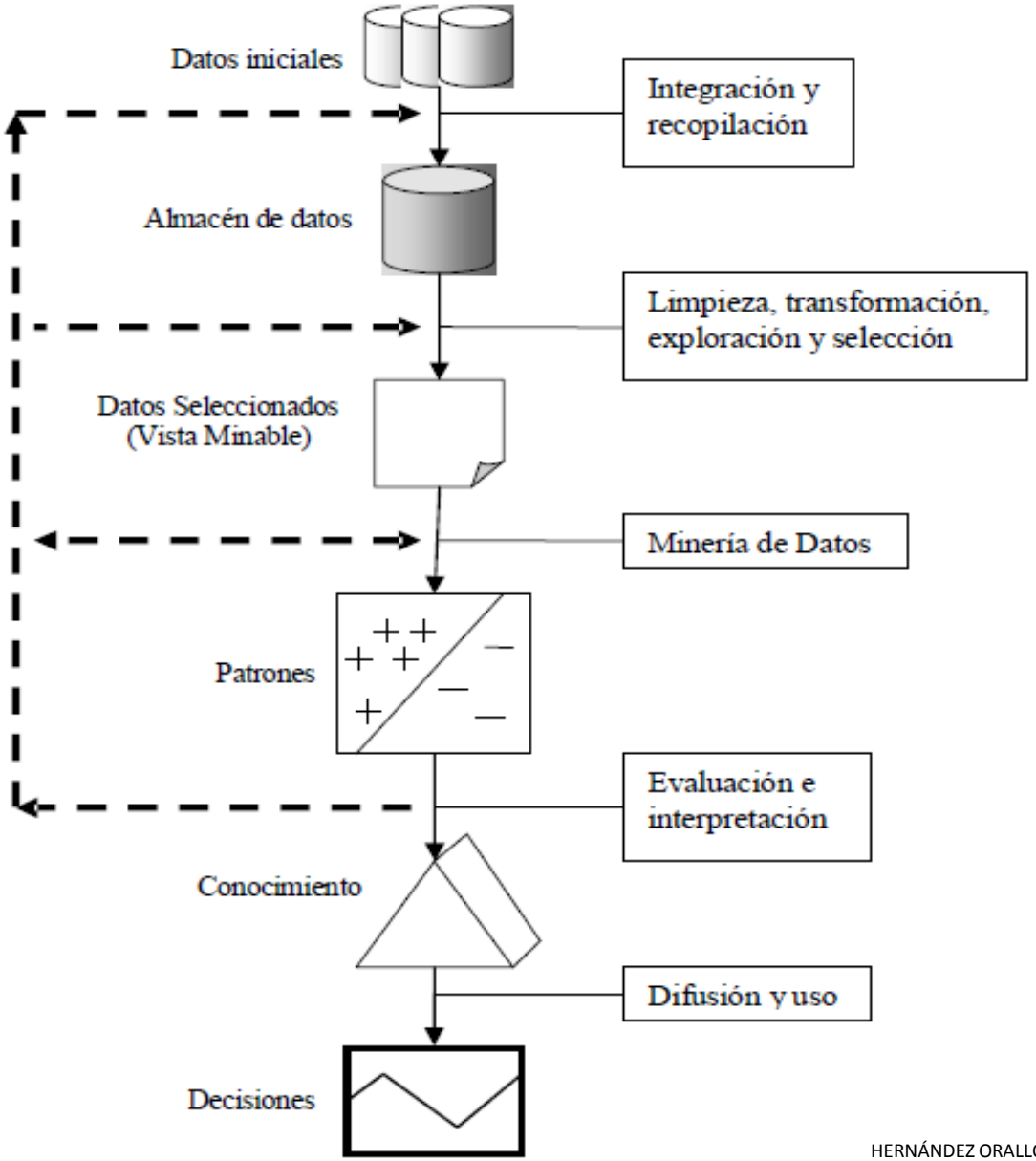

P3TQ
CRISP-DM
MoPro PEI
?
KDD
SEMMA

Proceso de KDD

Interactivo

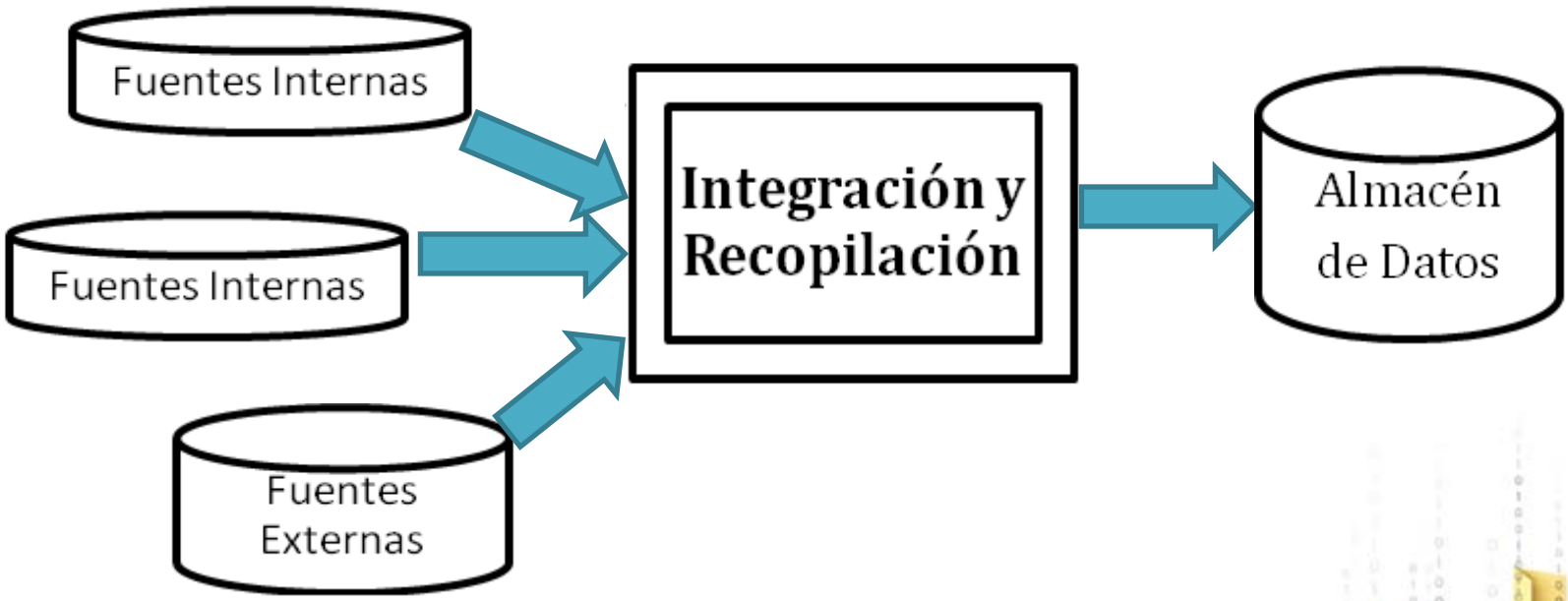
e

Iterativo



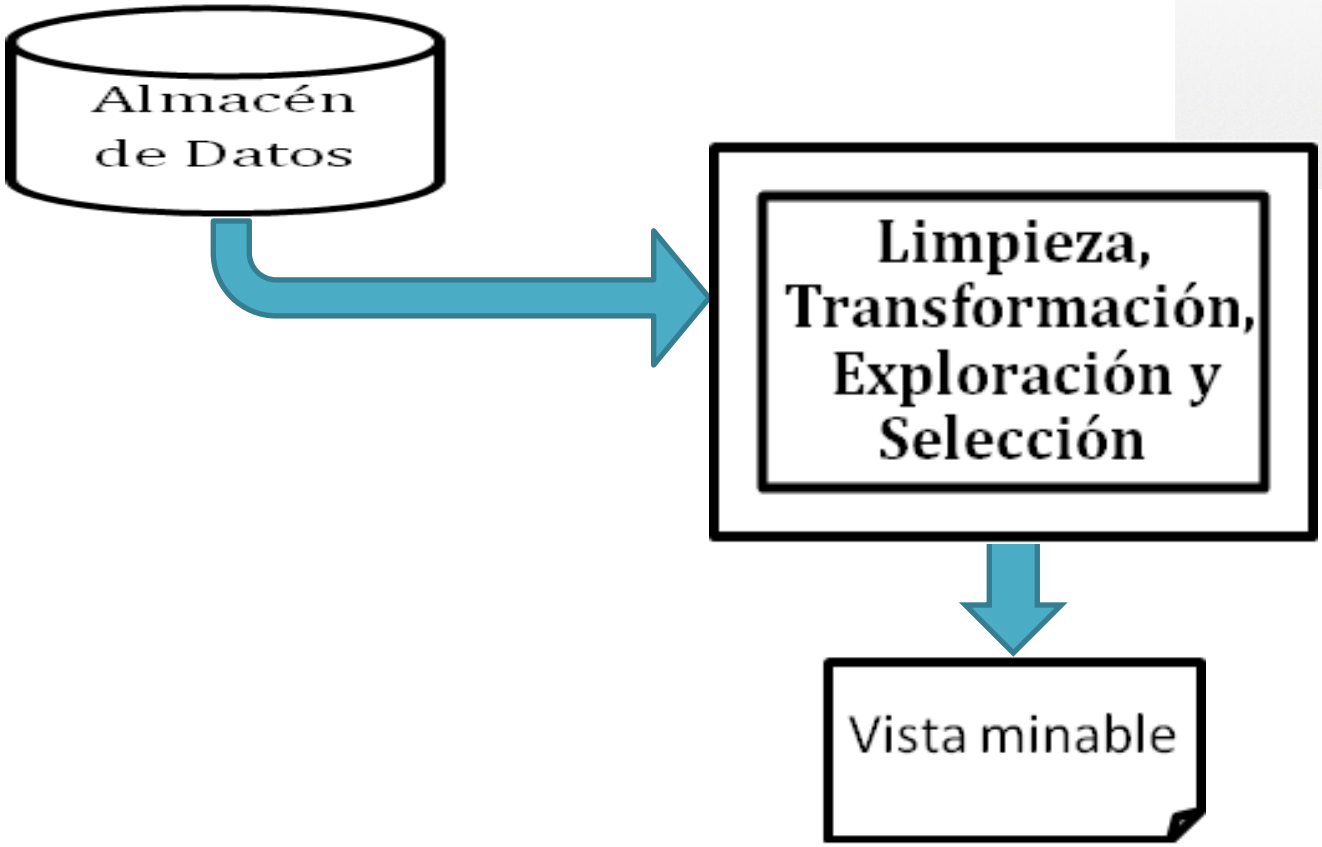
HERNÁNDEZ ORALLO, J. y otros. 2004

1. Integración y Recopilación



- Datos o fuentes de información en bruto.

2. Filtrado Datos

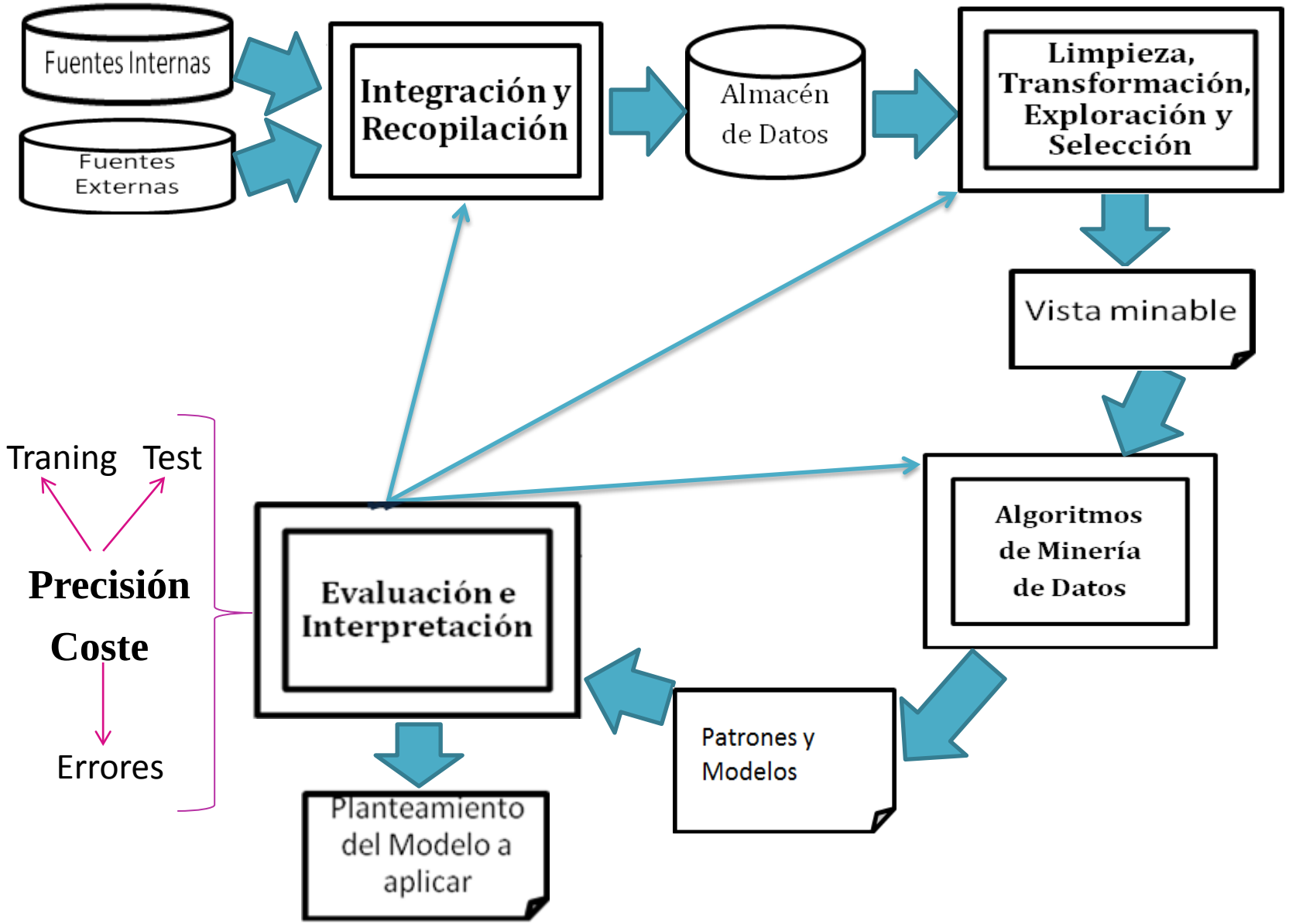


- Eliminar datos redundantes y filtrar los de mejor calidad para el proceso de minado.

3. Minería de Datos

- Crear un *Modelo* aplicando Técnicas y Algoritmos de MD para extraer *Patrones*.

4. Interpretación y Evaluación



5. Difusión, Uso y Monitorización

Finalidades:

- ***Acciones basadas en la observación del modelo y sus resultados.***
- ***Aplicación del modelo a diferentes conjuntos de datos.***

- Difusión y aplicación del modelo
  - Integrar al *know-how* de la organización.

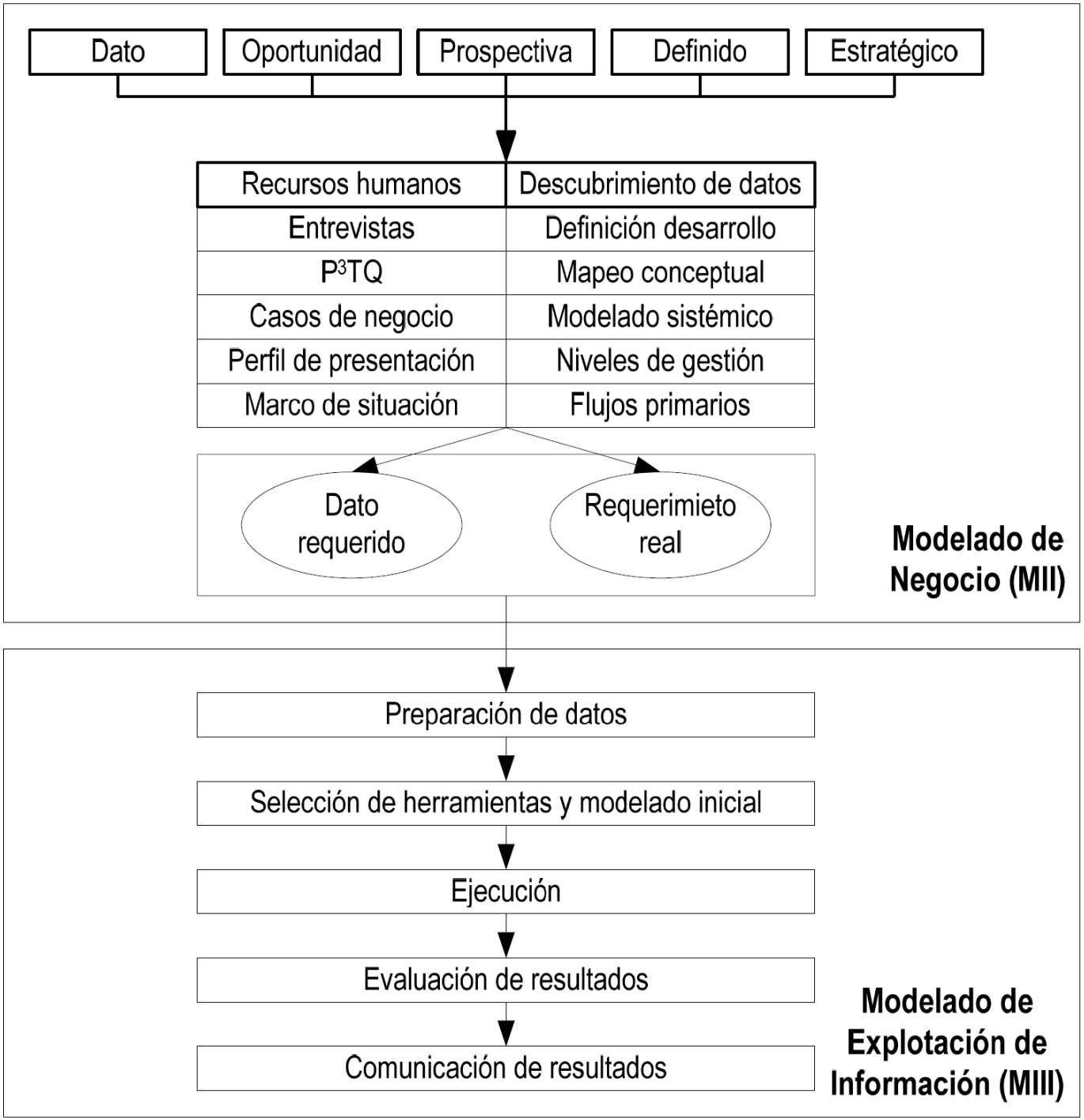
- Monitoreo y Mantenimiento.

Metodologías de Ciencia de Datos

Metodología indica **cómo encarar un Proy: fases, tareas y como llevarlas a cabo**

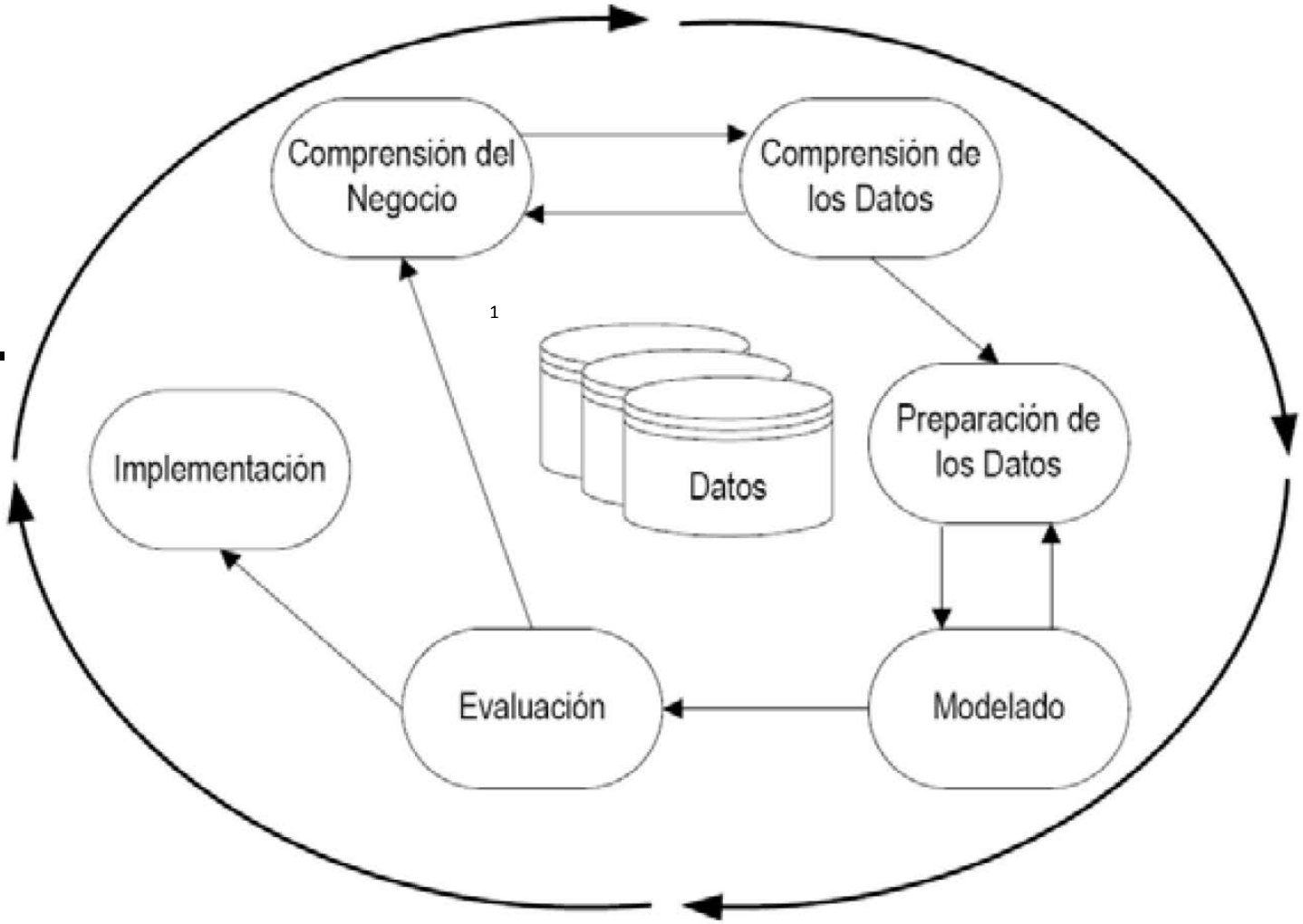
Catalyst P³TQ

(Product, Place, Price, Time, Quantity)



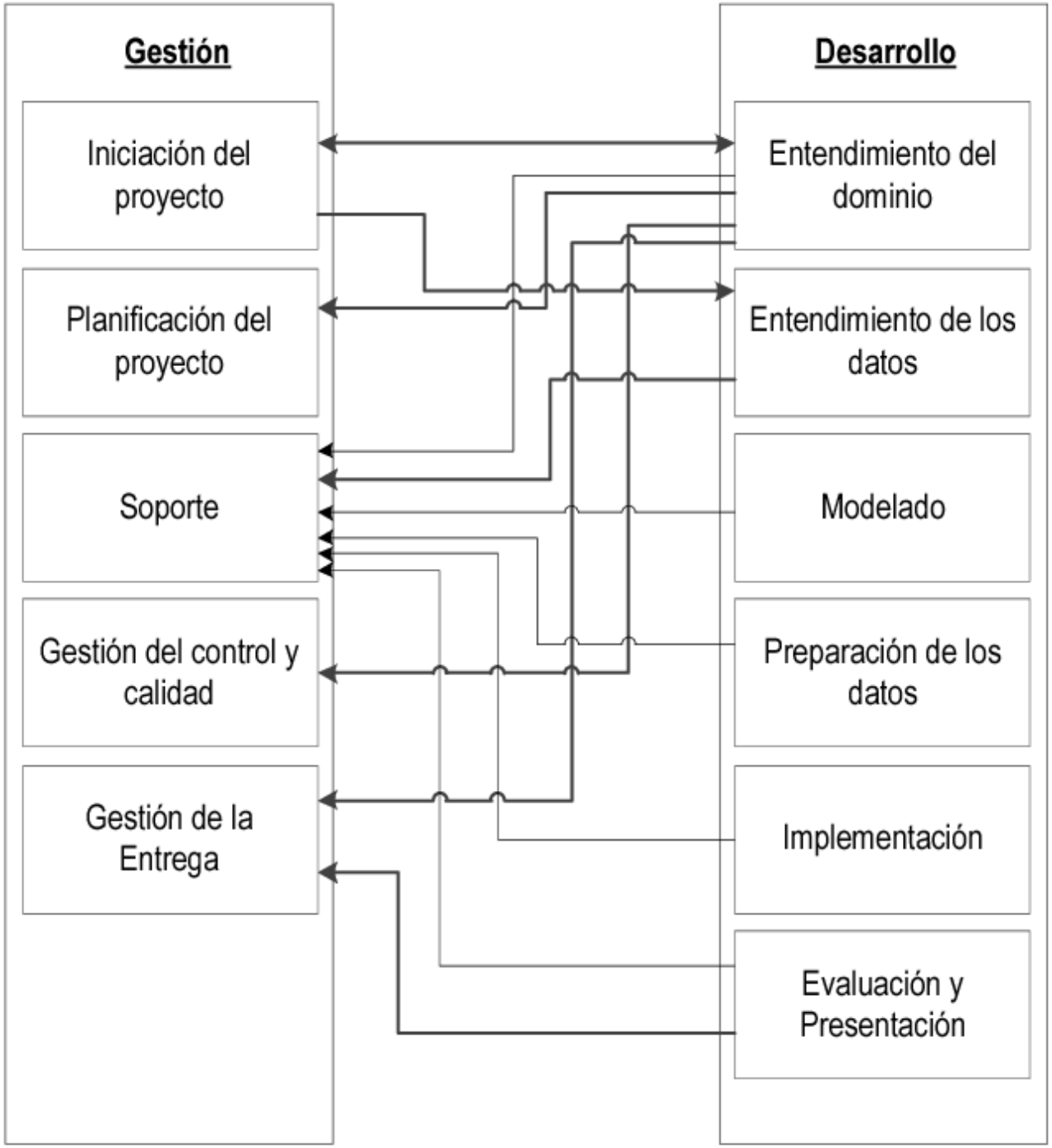
Pyle D. 2003

CRISP-DM



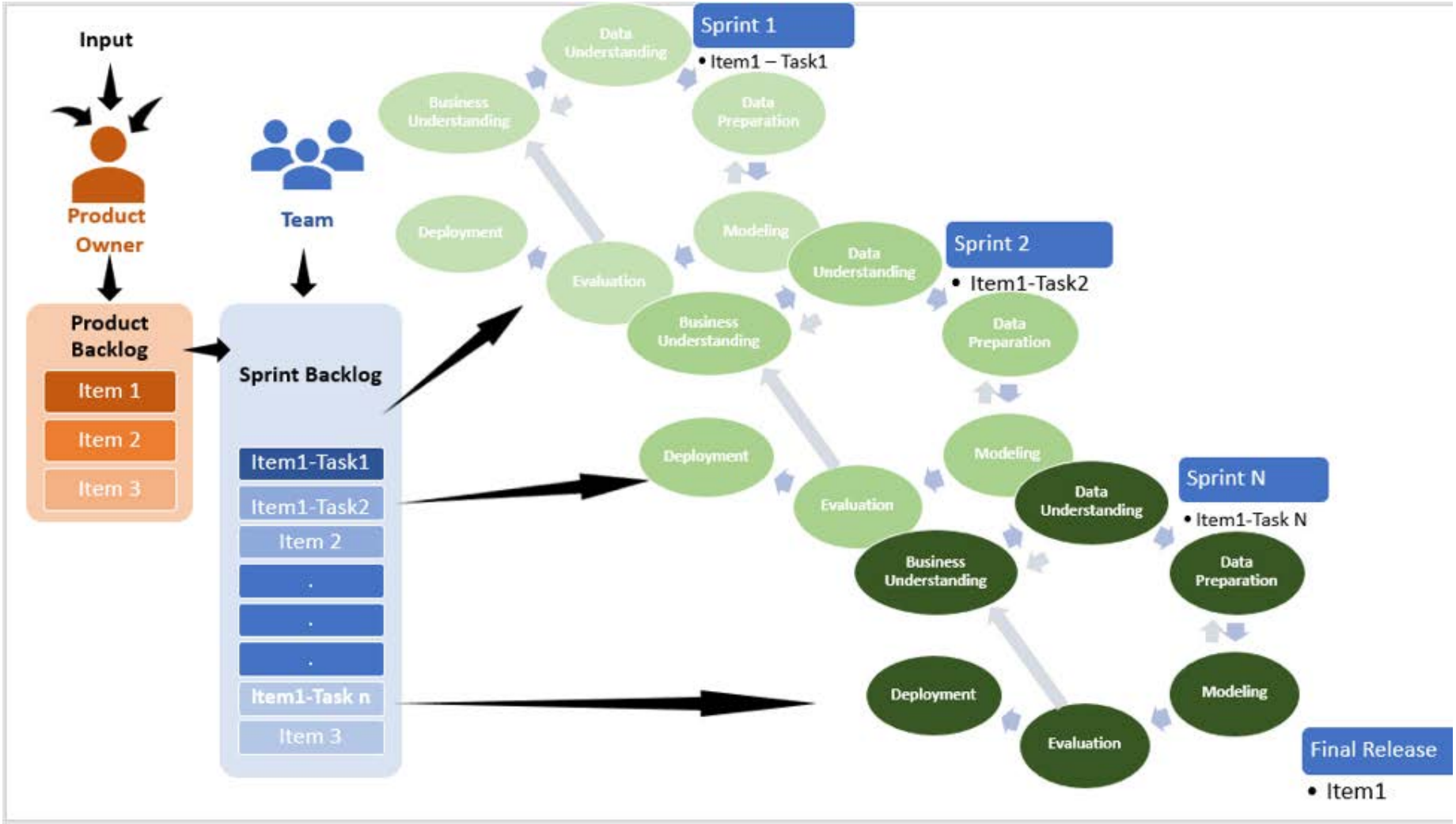
Chapman P. y otros. 2000

MoProPEI



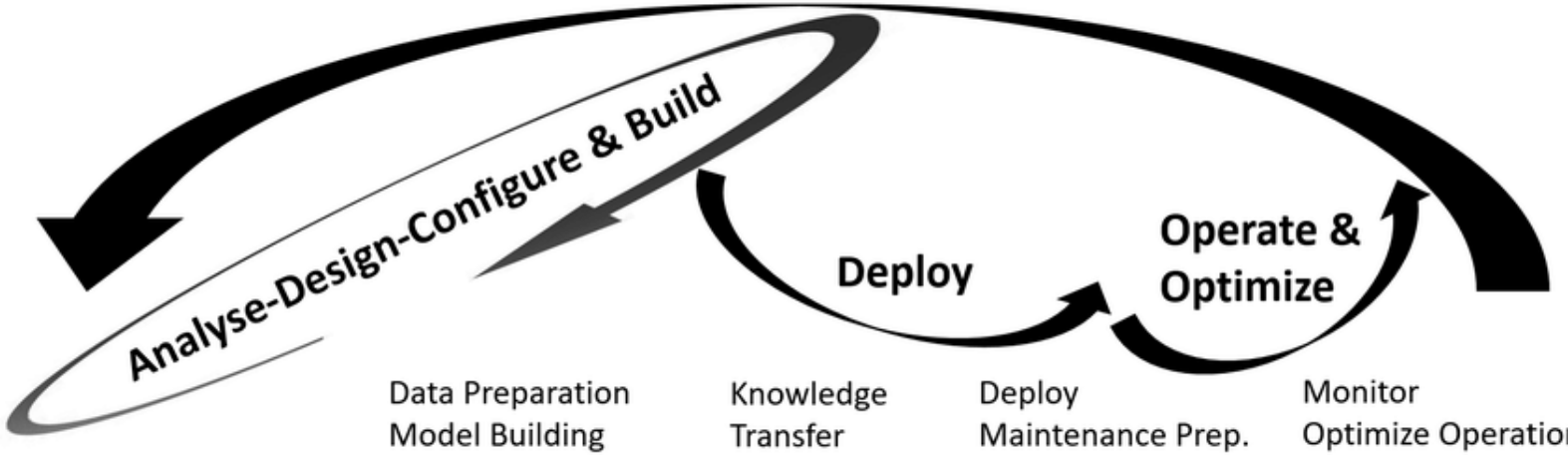
Martins S. y otros. 2014

Scrum-DM



Nameeta Raj , "CRISP-DM the Scrum Agile way. Why not!", 2018

ASUM-DM



Readiness check
Business Understanding
Data Understanding

Data Preparation
Model Building
Model Evaluation
Deployment Definition
Test & Operation Strategy

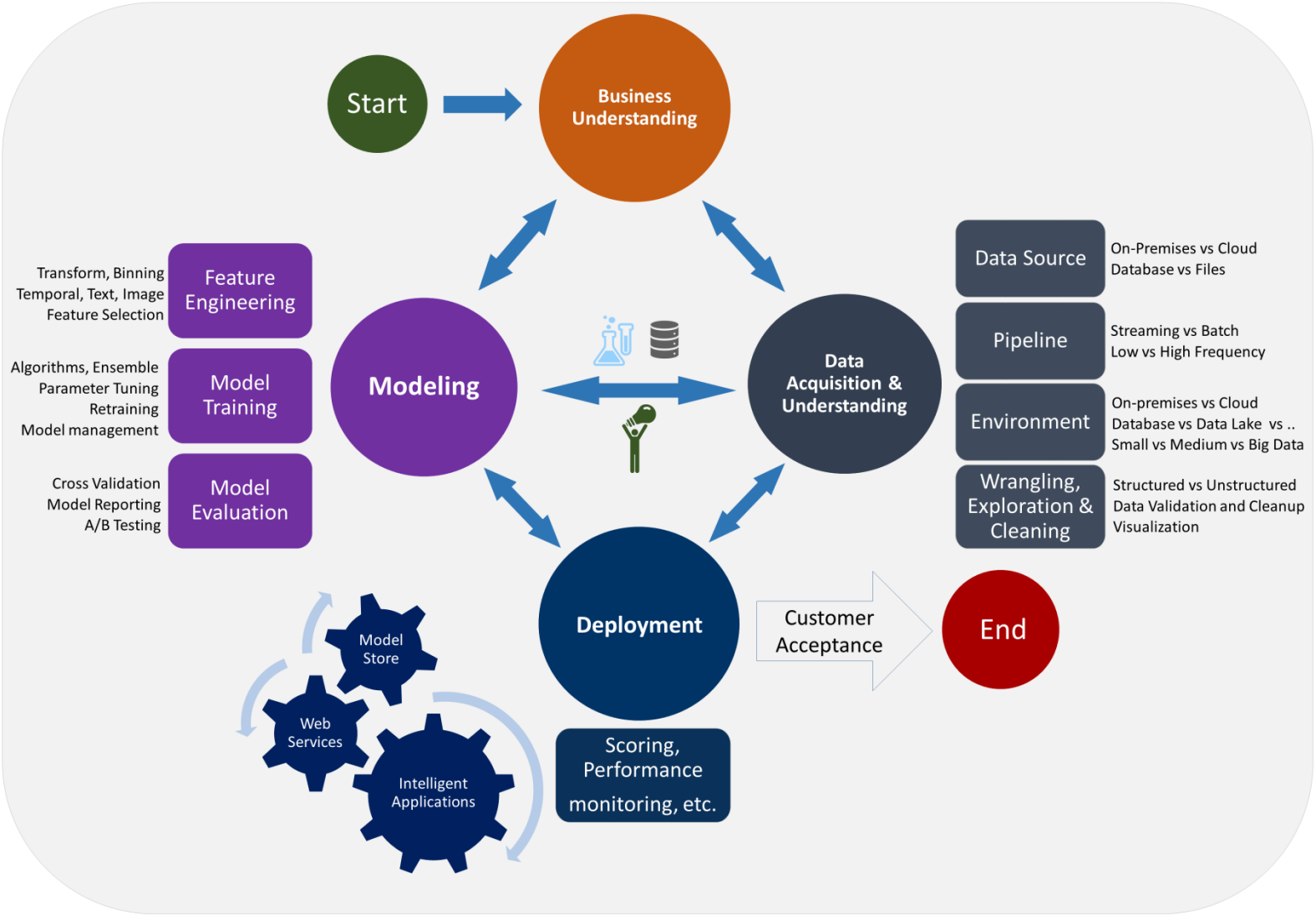
Knowledge Transfer
QA Validation

Deploy
Maintenance Prep.
Transfer to Support
Launch

Monitor
Optimize Operation
Support Users
Governance

Team Data Science Process (TDSP)

Data Science Lifecycle



LOS PROYECTOS DE DS NECESITAN UN PLAN DE ATAQUE CLARO Y EFICAZ PARA TENER **ÉXITO**

- **Comunicación,** para mostrar efectivamente los beneficios a los ejecutivos enseñando los resultados que se relacionan con los objetivos de la organización.

- **Entendimiento de negocio**, que solo ocurre a través de la interacción con las partes interesadas del negocio que están más cerca del proceso o problema.

- **Planificar y alinear** a todos los involucrados con el alcance y el plan del proyecto.

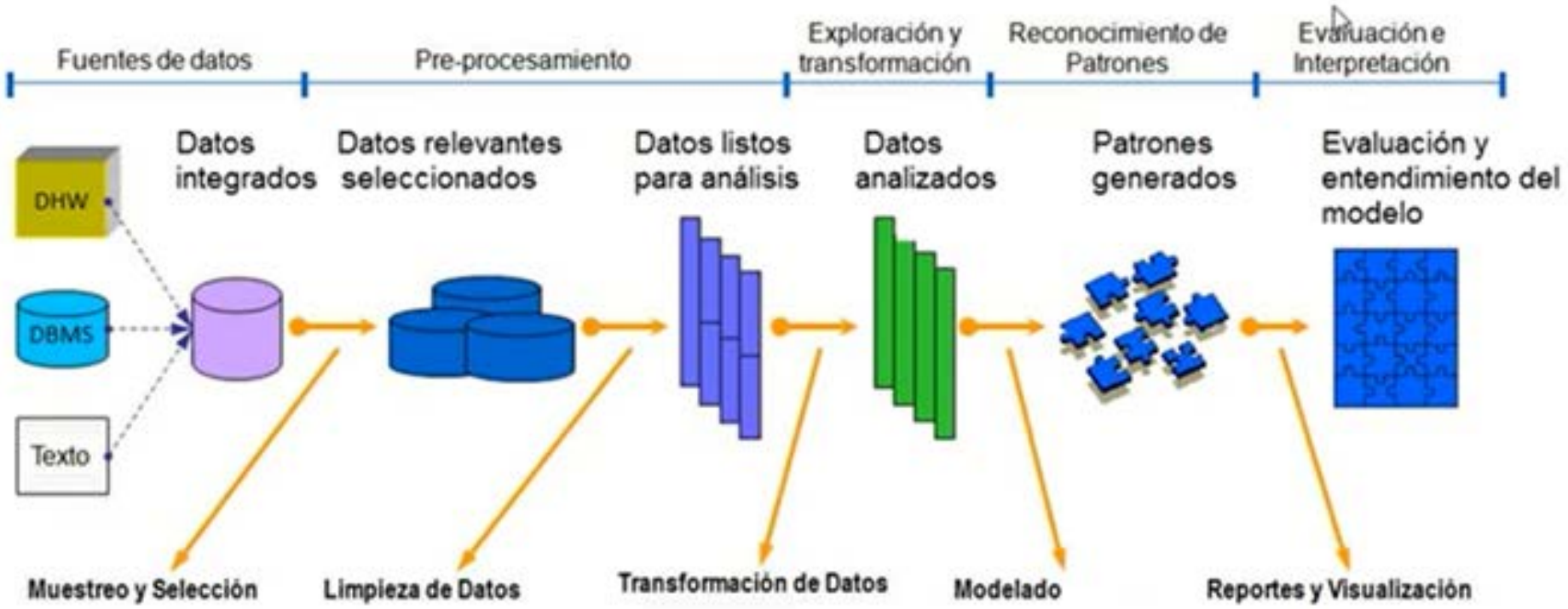
- Una **lista de acciones comprobadas** que deben considerarse.

En términos de datos

Una vez realizado el análisis del negocio y definido una estrategia, se plantean los objetivos técnicos que preguntas/conocimiento se deben responder/descubrir a partir de qué datos.

Procesamiento de datos

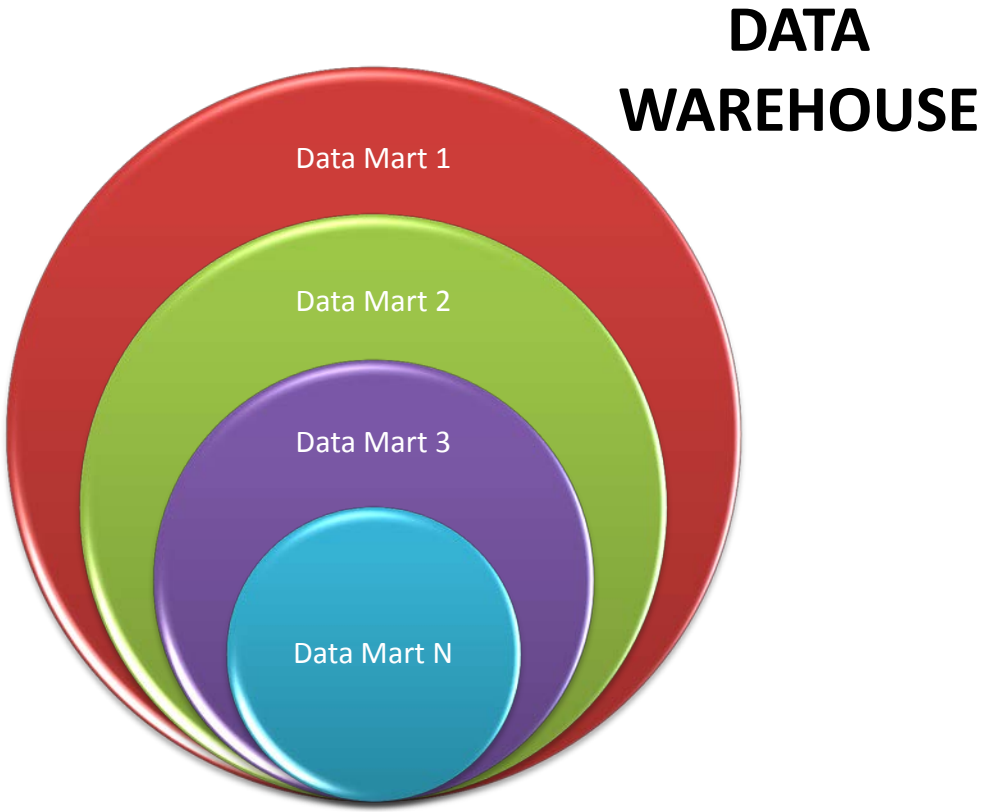
- Preparación y curado de los datos que permitan el modelado y tratamiento de éstos



Datos

Data Mart

Los Data Mart son subconjuntos de los DW, es decir que contienen subconjuntos de datos de toda la organización que son valiosos para diferentes grupos específicos de personas.



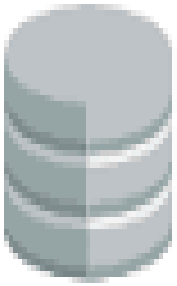
Datos

Data Warehouse

Un DW es un gran repositorio lógico de datos que permite el acceso y manipulación flexible de grandes volúmenes de información provenientes tanto de transacciones detalladas como datos agregados de fuentes de distintas naturaleza (Archivos planos, CSV, planillas de cálculo, etc).

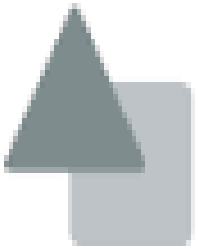
Un DW proporciona datos generalizados y consolidados en una vista multidimensional. Junto con una vista generalizada y consolidada de datos, un almacén de datos también nos proporciona herramientas de procesamiento analítico en línea (OLAP).

Datos

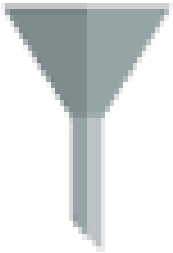


Data
Warehouse

colección de datos
para el soporte de la toma de decisiones



orientada
al negocio

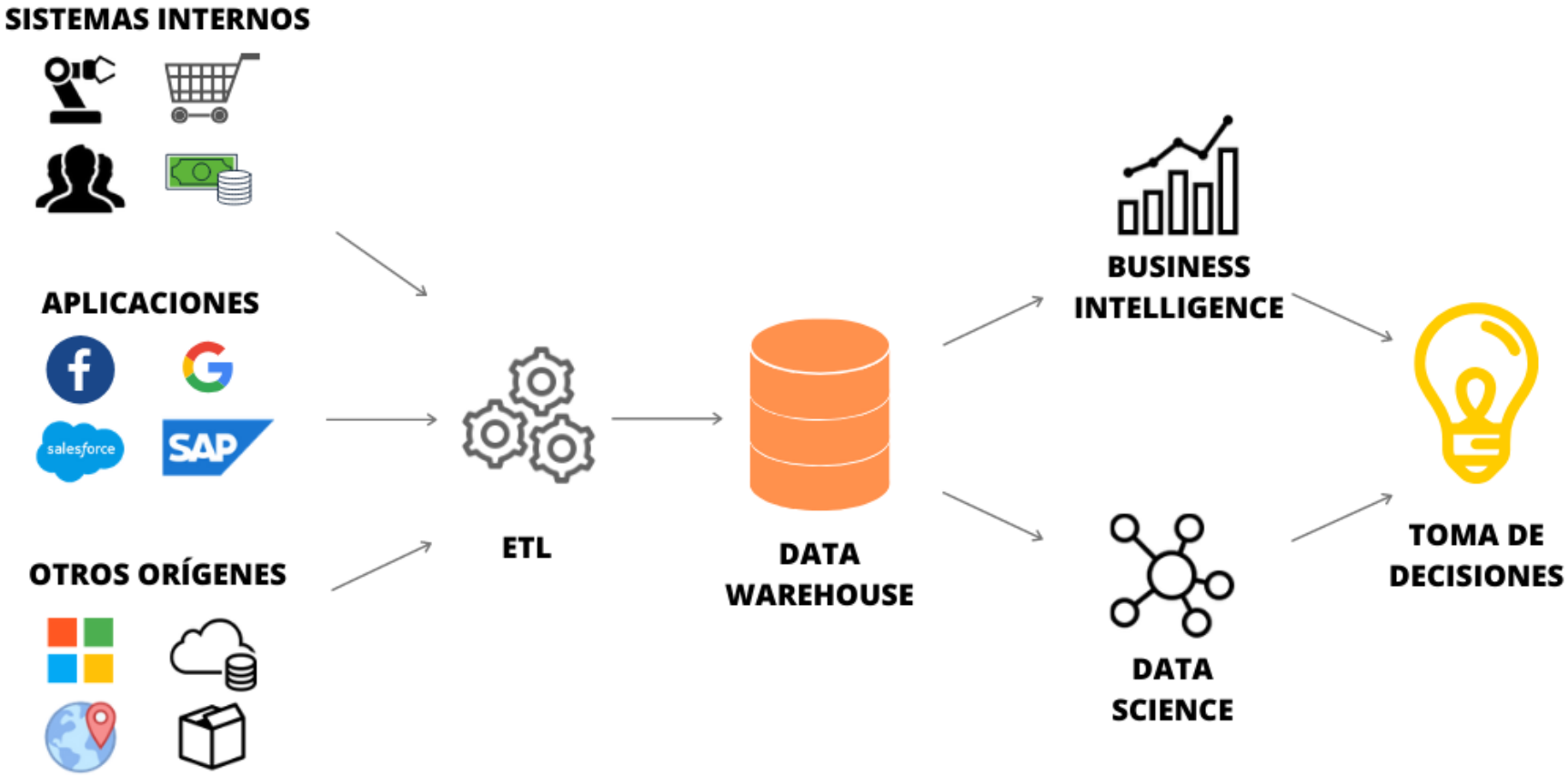


integrada

variante en
el tiempo

no volátil

Datos

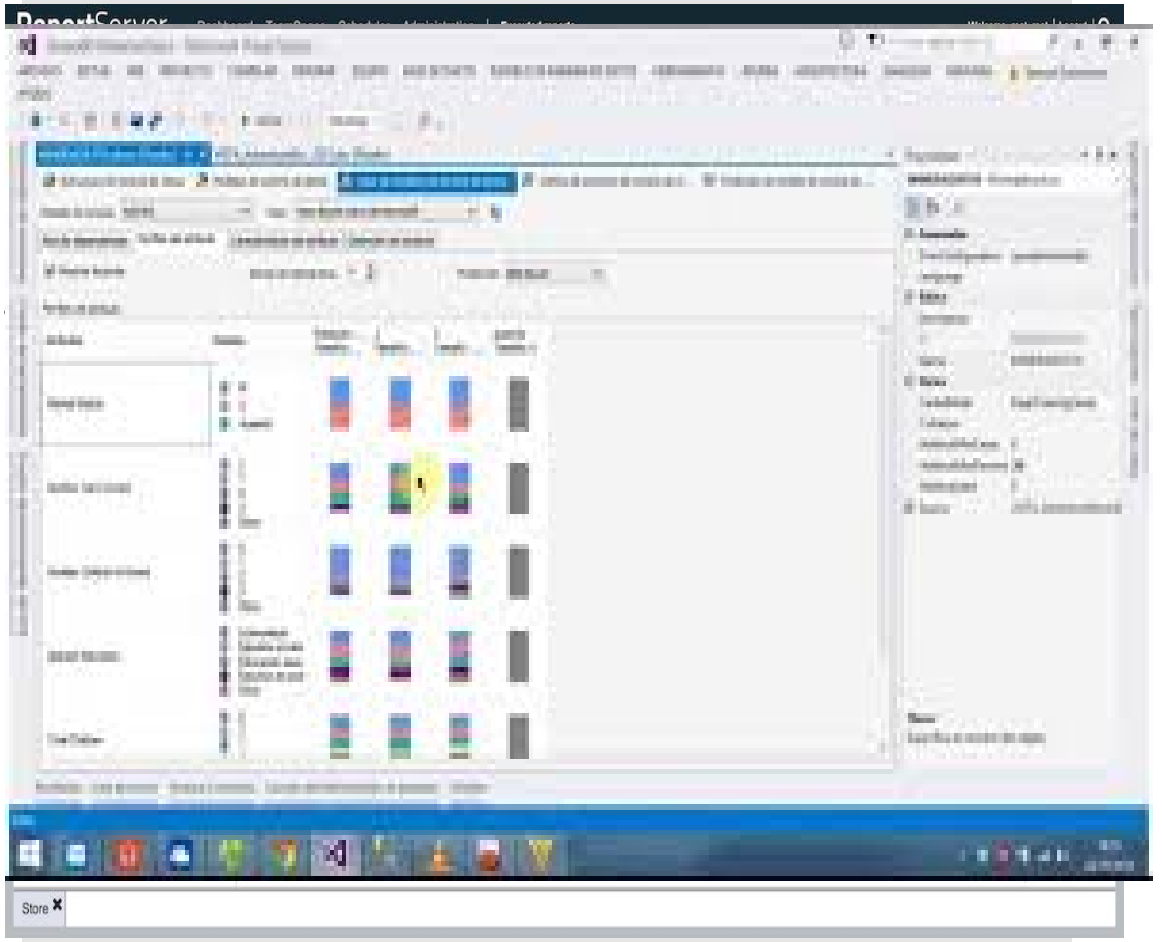


https://www.mistralbs.com/definicion-arquitectura-data-warehouse/

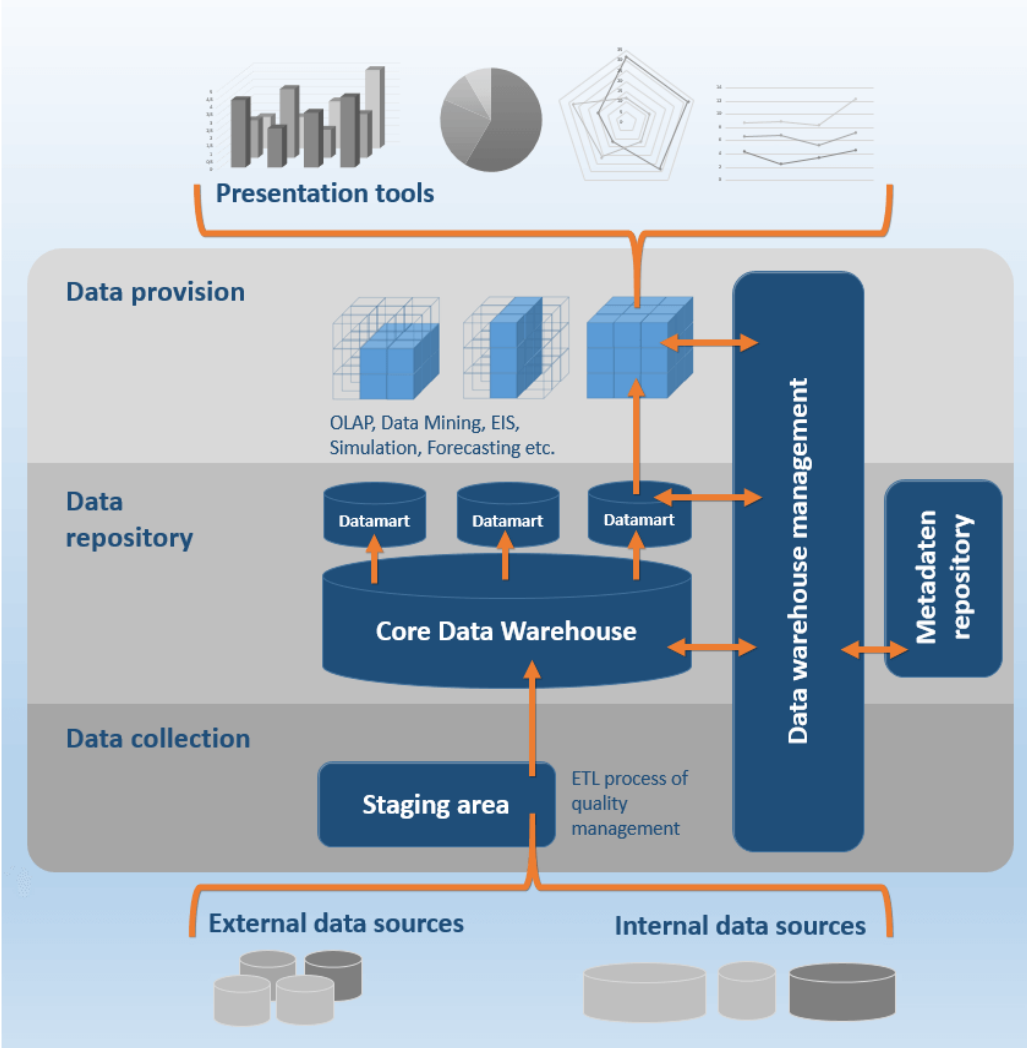
Datos

TIPOS DE DW

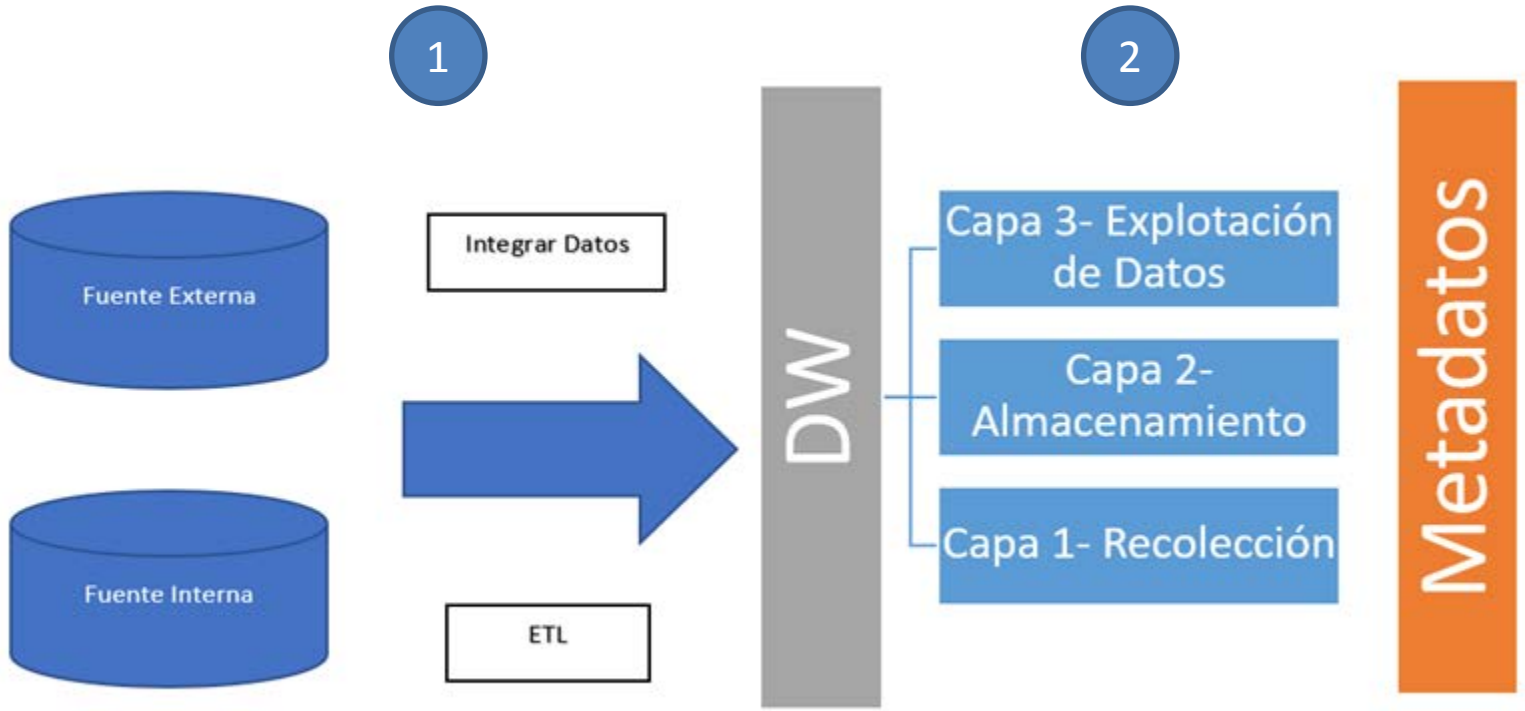
- Procesamiento de información.
- Procesamiento OLAP.
- Minería de Datos.



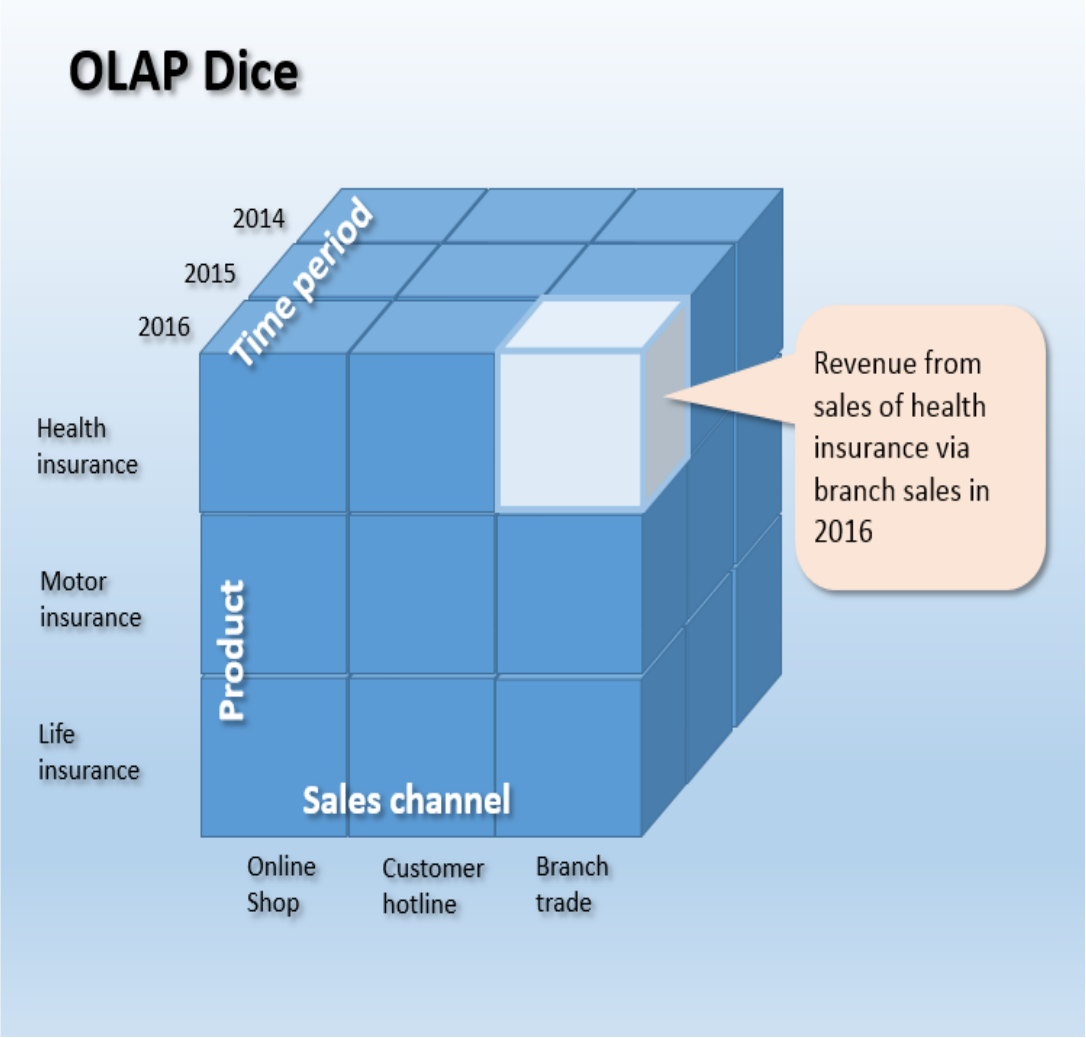
DATA WAREHOUSE – ARQUITECTURA



DATA WAREHOUSE – ARQUITECTURA



DATA WAREHOUSE – Cubo de Datos



OLAP vs MD

OLAP	Minería
¿Cuál es el promedio de accidentes entre los fumadores y los no fumadores?	¿Cuáles son los mejores vaticinadores para los accidentes?
¿Cuál es el promedio de la factura de teléfono de mis actuales clientes vs. mis exclientes?	¿Dejará X la compañía? ¿Qué factores afectan a las dimisiones?
¿Cuál es el promedio de compras diarias entre los usuarios de tarjetas de crédito robadas y usuarios legítimos?	¿Qué patrones están asociados al uso de tarjetas de crédito fraudulentas?

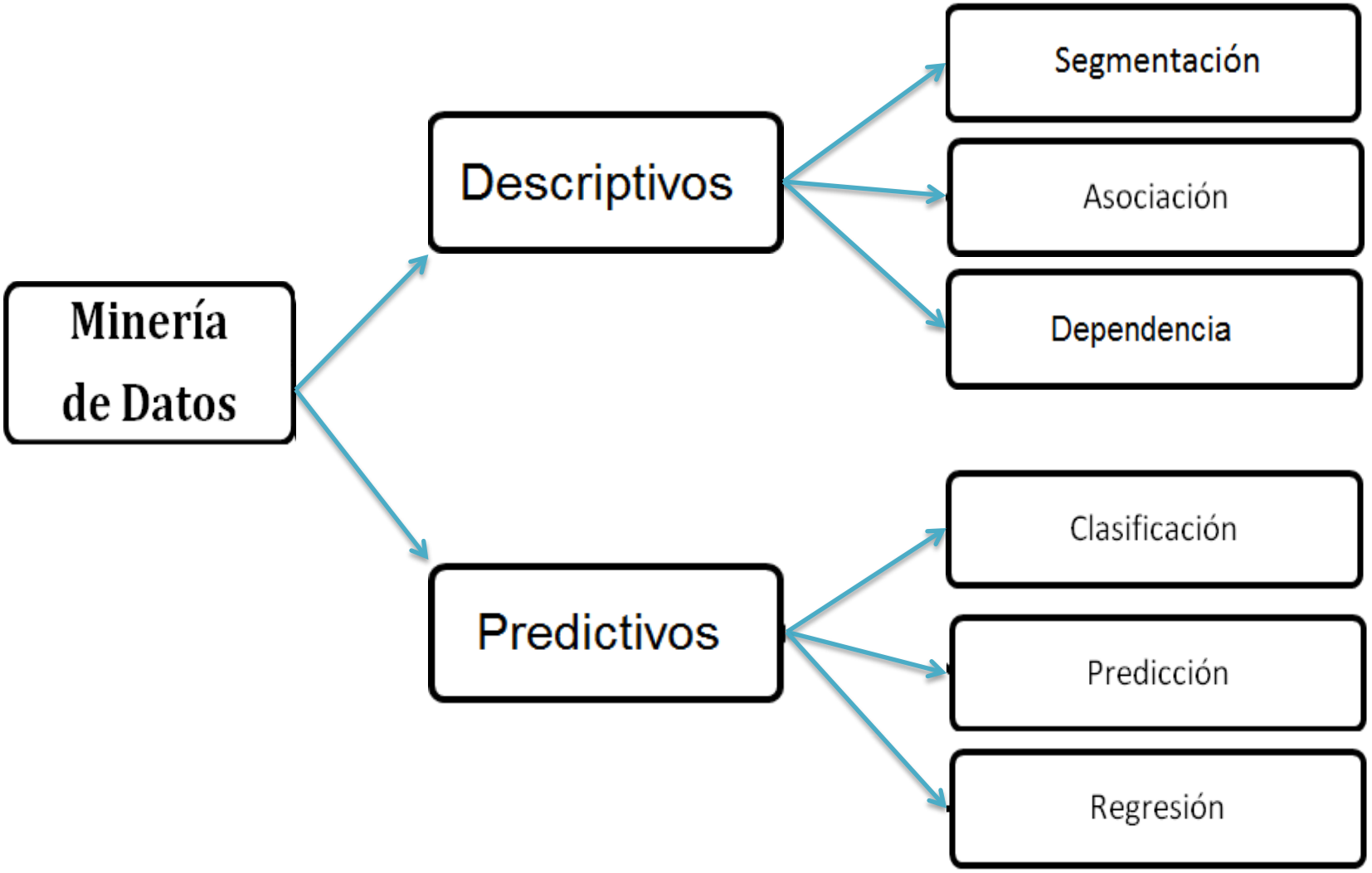
Datos

- Mis datos son valiosos para mi (in → in).
  - Datos internos útiles para la organización.
  - Inteligencia empresarial clásica... Muchas oportunidades todavía.
- Esos datos son valiosos para mi (out → in).
  - Datos externos útiles para la organización.
  - Medios sociales, Internet, datos abiertos, ... Muchas oportunidades nuevas.
- Mis datos son valiosos para otros (in → out).
  - Datos internos útiles para otras organizaciones.
  - Mis datos tienen utilidad para otros, ... Muchas oportunidades nuevas.
- Esos datos son valiosos para otros (out → out).
  - Datos externos útiles para otras organizaciones.
  - Estos datos tienen utilidad para otros, ... ¡Científico de datos *freelancer*!
- Creando datos (∅ → out).
  - Coleccionar datos que pueden tener valor. ¡Emprendedor de datos!

Datos

- Datos de telecomunicaciones
  - Valioso para comerciantes, tráfico, ayuntamiento, policía...
- Otros datos de geolocalización(*Flickr*, Instagram, *Wikiloc*, ...)
  - Valioso para agencias de viaje...
- Datos en consumo de energía
  - Valioso para anuncios de televisión...
- Datos del transporte público (bus, metro, tren, taxi, tráfico, ...)
  - Valioso para turismo, consumo, contaminación, comercio...
- Datos de redes sociales.
  - Valioso para casi todo...
- Datos de uso de tarjetas de crédito
  - Valioso para comercios, ayuntamientos, ...
- Datos de policía
  - Valioso para aseguradoras, agentes inmobiliarios, ...
- Datos comerciales (*Amazon*, *Ebay*, *segundamano.es*, ...)
  - Valioso para salud, demografía, sociología...
- Datos climatológicos
  - Valioso para comercios.
- Datos de búsquedas web.
  - Valioso para casi todo.

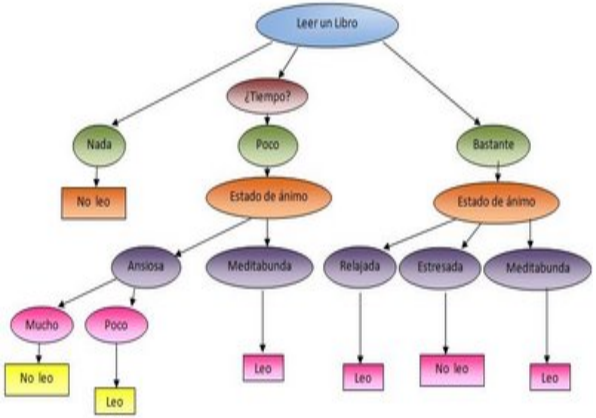
Modelos y Tareas

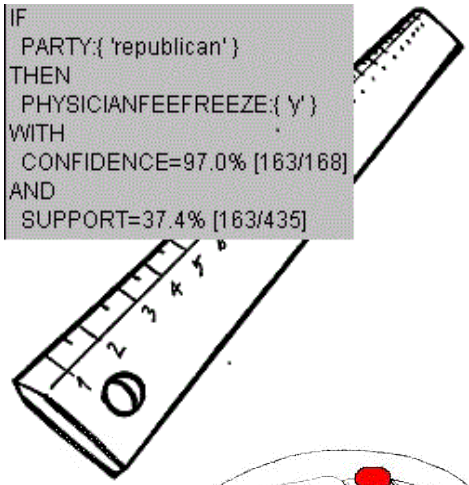


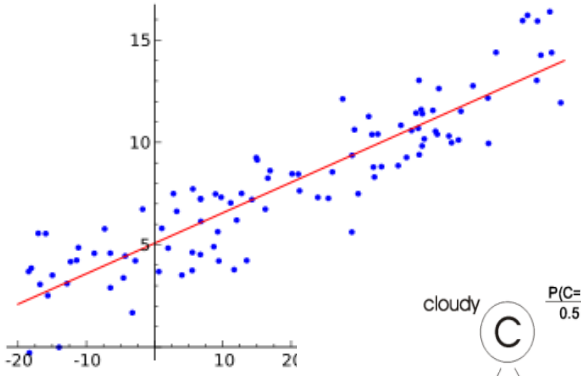
Tareas y Técnicas de MD

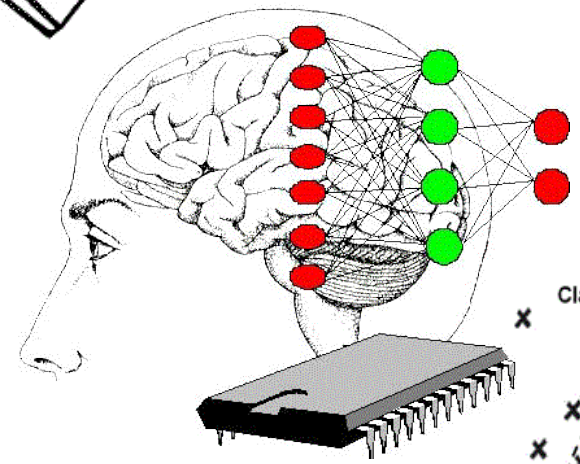
Descripción	Técnicas estadísticas (media, moda, mediana, desviación estándar, mínimo , máximo, rango, correlaciones) y gráficas, algoritmos genéticos
Clasificación	Redes neuronales (back propagation), árboles de decisión (ID3, C4.5, C5.0, CART), k-nn (k vecinos más cercanos), naive bayes, técnicas estadísticas
Estimación	Técnicas estadísticas (regresión lineal simple, correlación, regresión múltiple), árboles de decisión, k-nn, redes neuronales
Predicción	Técnicas estadísticas, redes neuronales, árboles de decisión, k-nn, algoritmos genéticos
Agrupación (Clustering)	Jerárquico, K-nn, K-means, Red Kohonen, Fuzzy C-means
Asociación	Apriori (all, some, dynamic some), GRI, FP Grow

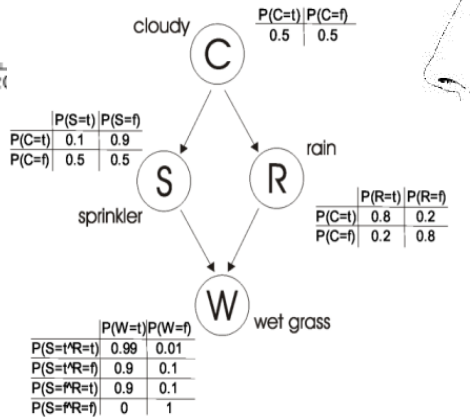
Algoritmos de MD

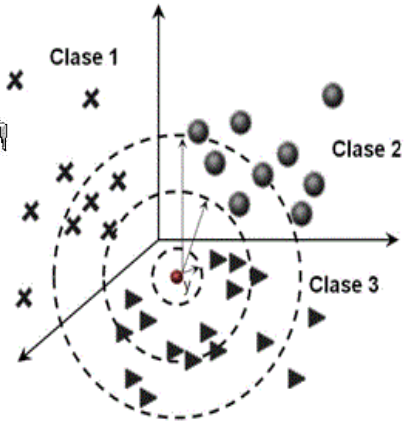










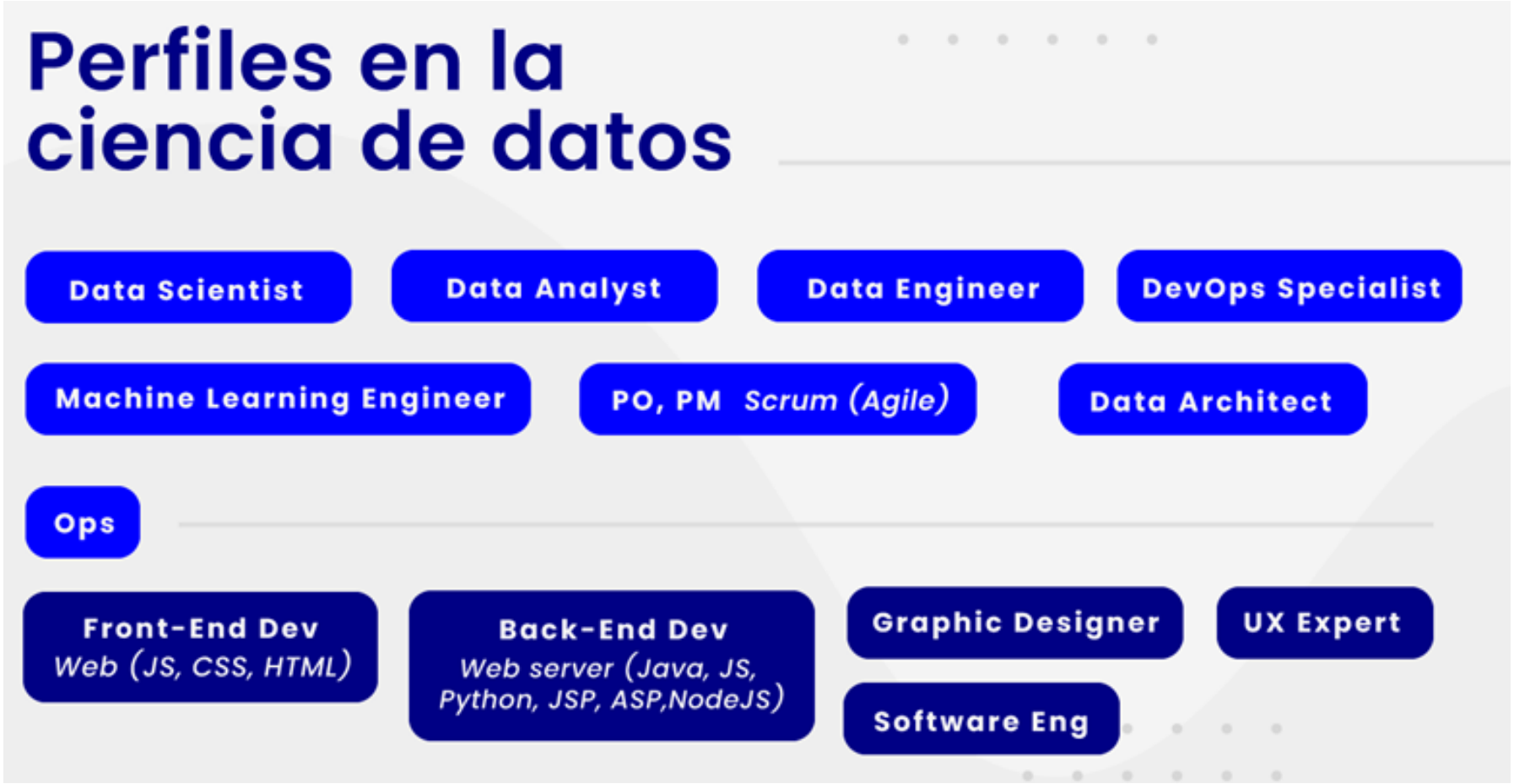


Kvecino más cercano

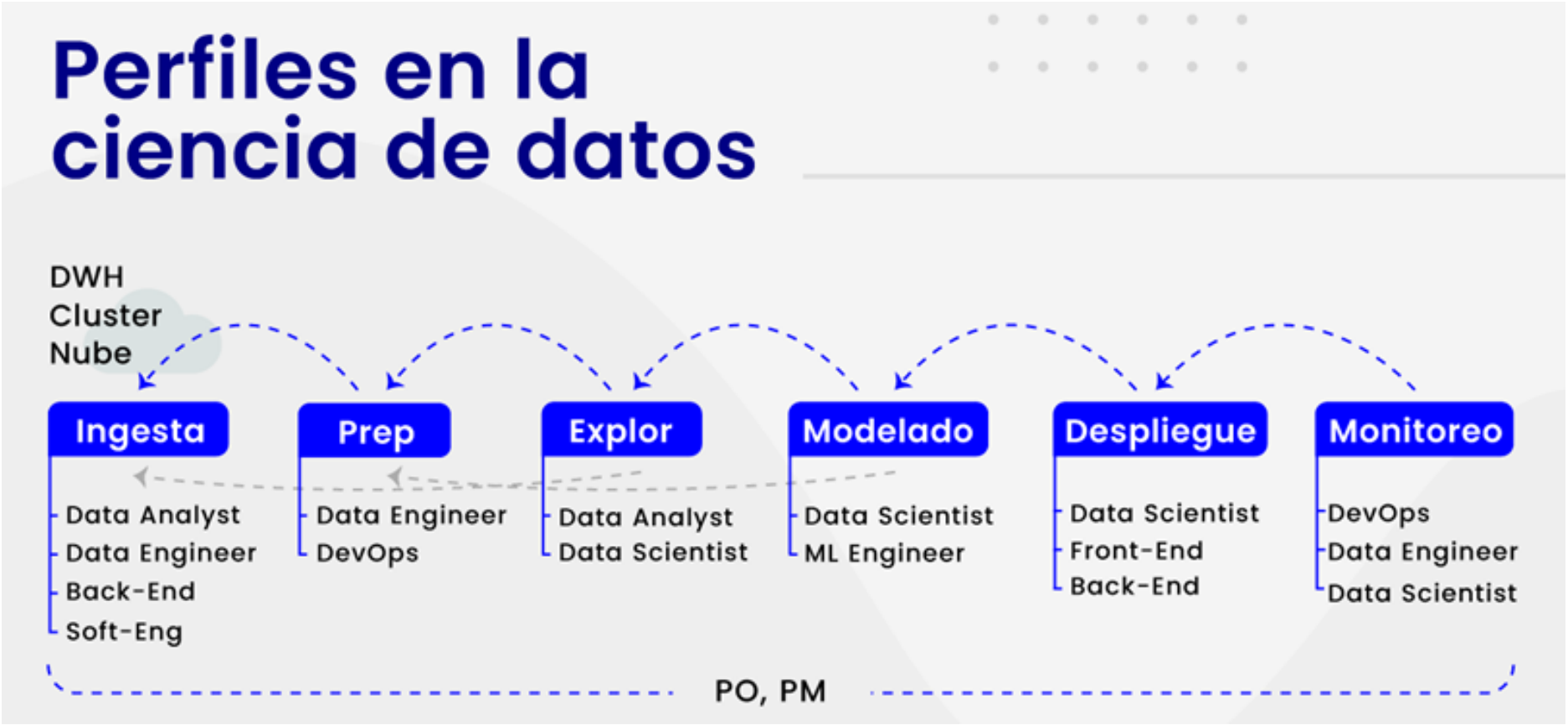
Modelos, Tareas y Técnicas de MD

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones/ Factorizaciones
Redes neuronales	✓	✓	✓		
Árboles de decisión ID3, C5.0	✓				
Árboles de decisiones CART	✓	✓			
Otros árboles de decisión	✓	✓	✓	✓	
Redes de Kohonen			✓		
Regresión lineal y logarítmica		✓			
Regresión logística	✓			✓	
Kmeans			✓		
Apriori				✓	
Naive Bayes	✓				
Vecinos más próximos	✓	✓	✓		
Análisis factorial y de componentes principales					✓
Twostep, Cobweb			✓		
Algoritmos genéticos y evolutivos	✓	✓	✓	✓	✓
Máquinas de vectores de soporte	✓	✓	✓		
CN2 rules (cobertura)	✓			✓	
Análisis discriminante multivariante	✓				

Ciencia de Datos



Ciencia de Datos



Ciencia de Datos

- Arquitecto de datos

Verifica recursos de infraestructura, personal, tipo de integración, tipo de análisis, cantidad de datos, tiempo de respuesta, tipo de información y legislaciones que se deben cumplir según el sector.

Determina qué sistemas de gestión de datos son apropiados según la estrategia de negocio.

Propone software, hardware, middleware que permitan la implementación de la solución.

Ciencia de Datos

- Ingeniero de datos

Los ingenieros de datos a menudo luchan con problemas asociados con la integración de bases de datos y conjuntos de datos no estructurados y desordenados. Su objetivo final es proporcionar datos limpios y utilizables a quien lo requiera.

El ingeniero debe tener una comprensión clara de cómo es el ciclo de vida de los datos y adecuarlos para reducir el componente de error humano.

Los ingenieros de datos limpian, preparan y optimizan los datos para el consumo.

Una vez que los datos se vuelven útiles se entregan a los científicos de datos.

Ciencia de Datos

- Científico de datos

Es una persona formada en ciencias matemáticas y computacionales con experiencia en cierta área de negocio o conocimiento que puede identificar que algoritmos y parámetros de análisis son los adecuados según la información con la que se cuenta para lograr ciertos objetivos.

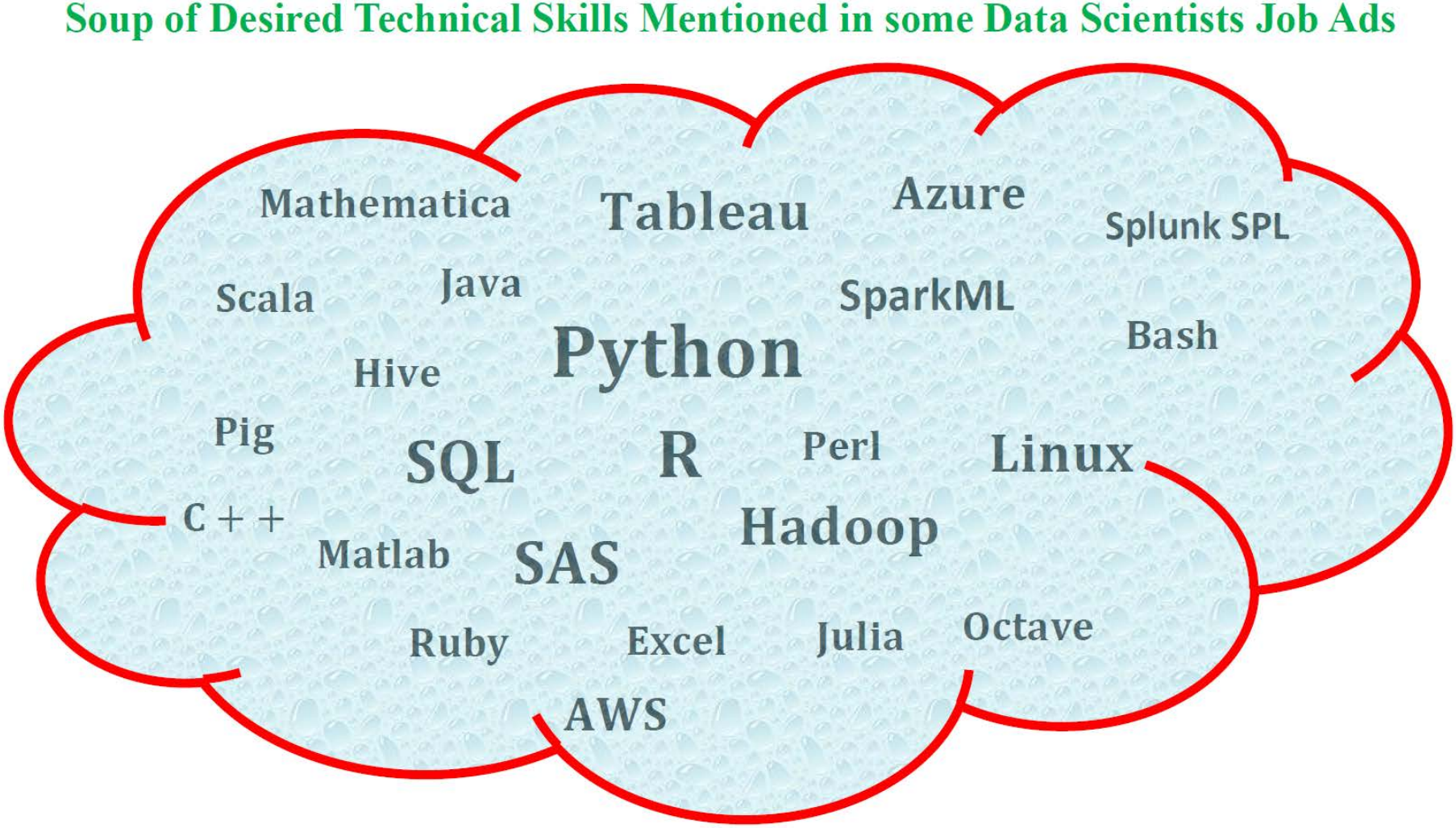
Además debe ser el enlace entre la estrategia de negocio, los métodos científicos, su interpretación y aplicación para lograr dichos objetivos.

Deben identificar que tarea es la mas conveniente y por cada tarea que técnica es la mejor, si no existe una, entonces diseñar y programar algoritmos que se ajusten a los datos y proporcionen el modelo matemático requerido.

Pueden realizar una variedad de análisis y técnicas de visualización para comprender verdaderamente los datos y, eventualmente, contar una historia, realizar predicciones o descubrir conocimiento a partir de los datos.

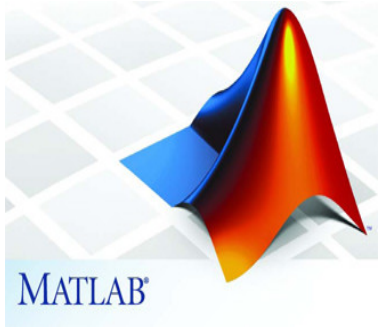
Ciencia de Datos

Lenguajes y Herramientas



Lenguajes de programación

&



Lenguajes de programación

Python

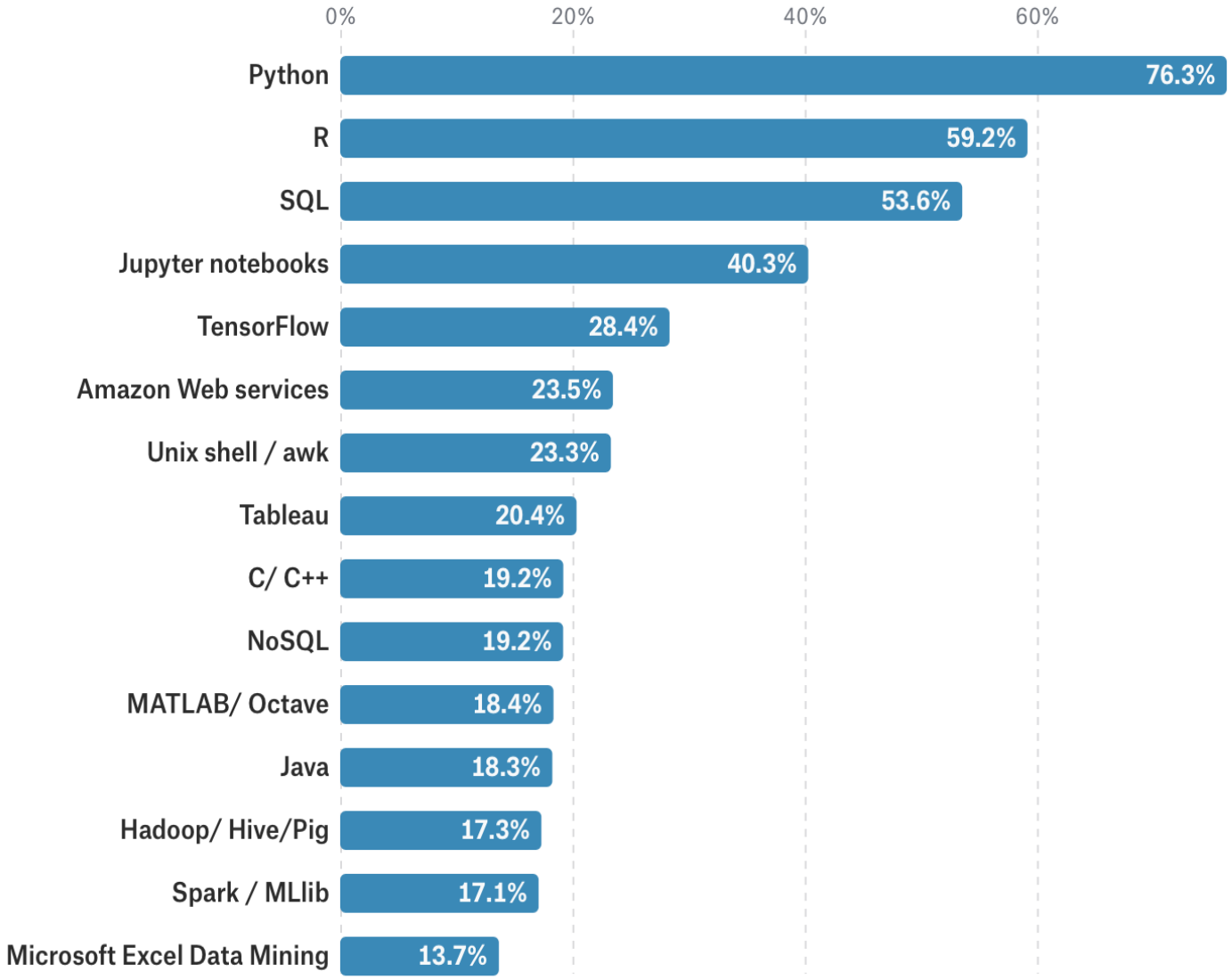
es un **lenguaje** de programación **interpretado**, orientado a **objetos** y de **alto nivel** con **semántica dinámica**.

Lenguajes de programación

R

es un **lenguaje** y **entorno para computación** y **gráficos estadísticos**

Lenguajes de programación



Herramientas de MD

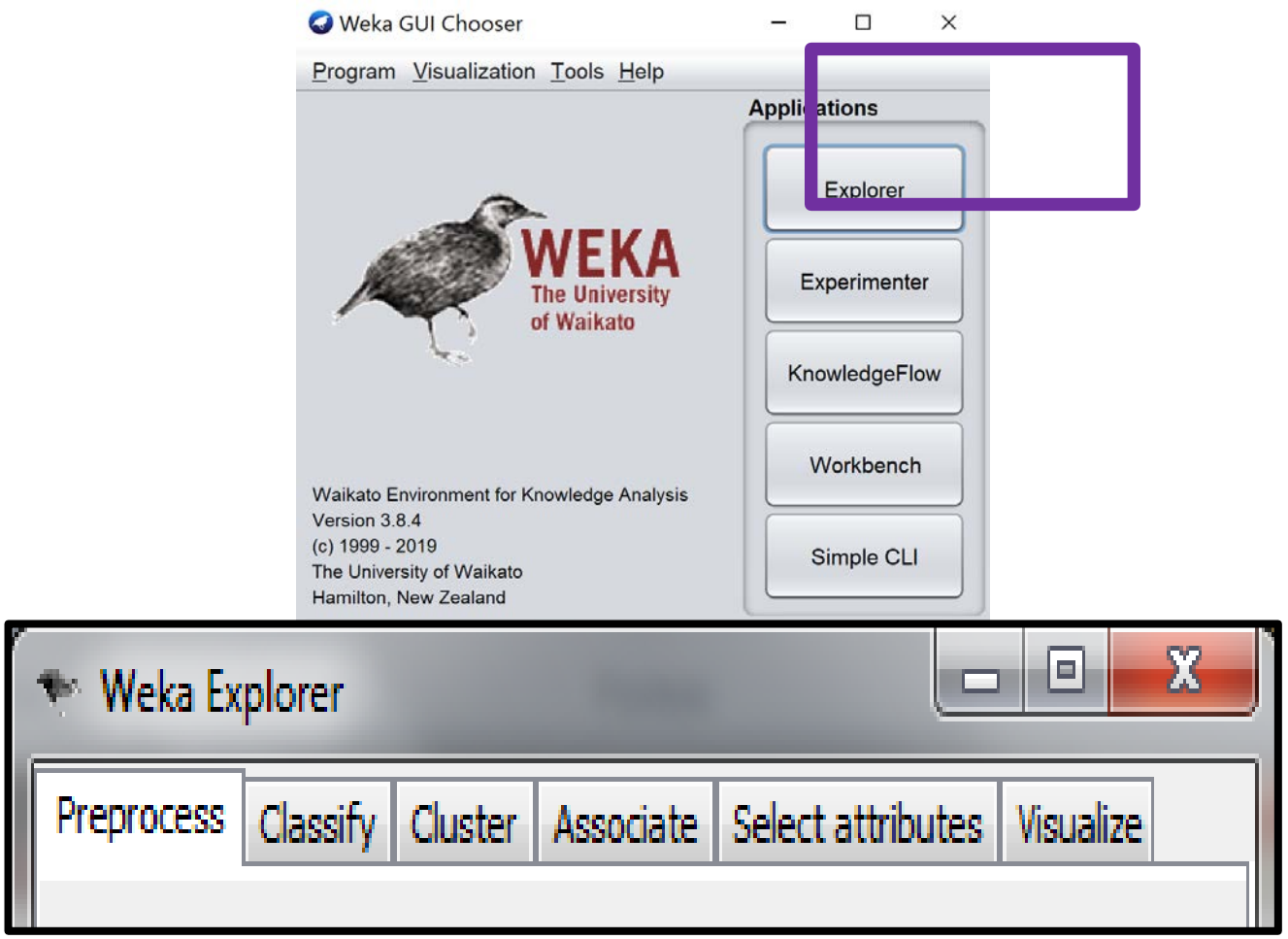
WEKA

Waikato Environment for Knowledge Analysis

- ***"El Tiempo es oro",*** es necesario aprovechar al máximo todos los recursos disponibles con la mayor velocidad posible.
- Implementada en Java.
- Licencia GPL.
- Específicamente diseñada y utilizada para investigación y fines educativos.

Tanagra 1.4

Interfaz Gráfica de Usuario



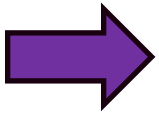
https://www.cs.waikato.ac.nz/ml/weka/

Aplicaciones

- Educación
- Agricultura
- Paso de fronteras
- Energía eléctrica
- BI
- PLN en diferentes ámbitos
- Proc. de imagen en diferentes ámbitos

Educación

- Problemas:
  - Calidad académica
  - Deserción
  - Reprobación
  - Retraso estudiantil
  - Bajos índices de eficiencia



Estrategias y Metodologías de anticipación a Eventos

EDM

Beneficios MDE

Objetivo Pedagógico

- Detección temprana de:
  - Casos críticos de deserción.
  - Retención de alumnos destacados.
  - Intervención de áreas con altos indicadores de reprobación.

Objetivo Comercial

- Mejorar la distribución de recursos :
  - Optimizar la adjudicación de becas y capacitaciones.
  - Adquisición bibliográfica.
- Detectar y potenciar intereses por carreras o cursos específicos.

Objetivos de Gestión

- Impresión objetiva, basada en el desempeño del alumno:
  - Promedio de calificaciones.
  - Uso de los recursos bibliográficos para su plan de estudio.
  - Participación en proyectos institucionales y en ayudantías.
- Informar anticipadamente sobre:
  - Comportamiento académico.
  - Capacitaciones especiales.
  - Deudas.

Ámbito Educativo

"Aplicación de Técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD"

Objetivo

- Determinar patrones y relaciones entre datos de la trayectoria académica de los alumnos de la UGD utilizando técnicas de Minería de Datos (MD).

  – Demás integrantes
    - Roberto Suénaga, Luciano Duarte (QEPD)

Preparación, Pruebas y Desarrollo

Herramientas Utilizadas

1. Integración y Recopilación de Datos

a) Análisis BD UGD

Tabla: Alumnos	Tabla: Aprobacion
Tabla: Asistencia Alumnos	Tabla: Asistencias Generadas
Tabla: Carrera por Alumno	Tabla: Carreras
Tabla: CLASE_TIPO	Tabla: Clases Especiales
Tabla: Clases por Materia	Tabla: Clases Regulares
Tabla: Comisiones	Tabla: Comisiones Habilitadas
Tabla: Condicion	Tabla: Condicion Cursado
Tabla: Condicion Examen	Tabla: Condicion Fichadas
Tabla: Condicion Regularidad	Tabla: Correlativas
Tabla: Cursado	Tabla: Documentacion Legajo
Tabla: Encabezado Parciales	Tabla: Equivalencia
Tabla: Establecimientos	Tabla: Estado Civil
Tabla: Estados Asistencia	Tabla: Historial Alumnos
Tabla: Informes Desempeño	Tabla: Inscripcion a Examen
Tabla: Legajo Alumno	Tabla: Localidad
Tabla: Materias	Tabla: Matriculaciones
Tabla: Matriculas	Tabla: Mesas
Tabla: Mesas de Examen	Tabla: Modalidad Carrera
Tabla: Modo Aprobacion	Tabla: Modo Informe
Tabla: Pais	Tabla: Parciales
Tabla: Profesores	Tabla: Provincia
Tabla: Regularidad	Tabla: Sede
Tabla: Tipo de Documento	Tabla: Tipo Parcial
Tabla: TipoDocumento	Tabla: Titulos
Tabla: Titulos Ciclo Profesorado	Tabla: Titulos de Grado
Tabla: Titulos Otorgados	Tabla: Turno Examen

b) Consolidación de los Datos Finales

_CONSULTA_TESIS_MD_UGD

```
SELECT DISTINCT Carreras.[Grupo Plan], Carreras.AreaDpto AS Dpto, Condicion.DESCRIPCION AS Condicion, UltimoAñoCarrCurs.MáxDeAño AS UltAñoAcadCurs, MatCursAñoAcadCarr.[1] AS 1°Curs,
MatAprAñoAcadCarr.[1] AS 1°Apro, MatCursAñoAcadCarr.[2] AS 2°Curs, MatAprAñoAcadCarr.[2] AS 2°Apro, MateriasCuatrimeAño.[1999] AS Curs99, MateriasCuatrimeAñoCondicion.[1999] AS FracCurs99,
MateriasRendAñoss.[1999] AS Apro99, MateriasCuatrimeAño.[2000] AS Curs00, MateriasCuatrimeAñoCondicion.[2000] AS FracCurs00, MateriasRendAñoss.[2000] AS Apro00, MateriasCuatrimeAño.[2001]
AS Curs01, MateriasCuatrimeAñoCondicion.[2001] AS FracCurs001, MateriasRendAñoss.[2001] AS Apro01, MateriasCuatrimeAño.[2002] AS Curs02, MateriasCuatrimeAñoCondicion.[2002] AS FracCurs02,
MateriasRendAñoss.[2002] AS Apro02, MateriasCuatrimeAño.[2003] AS Curs03, MateriasCuatrimeAñoCondicion.[2003] AS FracCurs03, MateriasRendAñoss.[2003] AS Apro03, MateriasCuatrimeAño.[2004]
AS Curs04, MateriasCuatrimeAñoCondicion.[2004] AS FracCurs04, MateriasRendAñoss.[2004] AS Apro04, MateriasCuatrimeAño.[2005] AS Curs05, MateriasCuatrimeAñoCondicion.[2005] AS FracCurs05,
MateriasRendAñoss.[2005] AS Apro05, MateriasCuatrimeAño.[2006] AS Curs06, MateriasCuatrimeAñoCondicion.[2006] AS FracCurs06, MateriasRendAñoss.[2006] AS Apro06, MateriasCuatrimeAño.[2007]
AS Curs07, MateriasCuatrimeAñoCondicion.[2007] AS FracCurs07, MateriasRendAñoss.[2007] AS Apro07, MateriasCuatrimeAño.[2008] AS Curs08, MateriasCuatrimeAñoCondicion.[2008] AS FracCusr08,
MateriasRendAñoss.[2008] AS Apro08, MateriasCuatrimeAño.[2009] AS Curs09, MateriasCuatrimeAñoCondicion.[2009] AS FracCusr09, MateriasRendAñoss.[2009] AS Apro09, MateriasCuatrimeAño.[2010]
AS Curs10, MateriasCuatrimeAñoCondicion.[2010] AS FracCurs10, MateriasRendAñoss.[2010] AS Apro10, PromAprob.[Total de Calificacion] AS PromAproC, PromAño.[Total de Calificacion] AS PromGralC,
PromAprob.[1] AS PromAp1, PromAño.[1] AS PromGr1, PromAprob.[2] AS PromAp2, PromAño.[2] AS PromGr2, DateDiff("yyyy",[Fecha Nacimiento],[Carrera por Alumnos].[Fecha Ingreso])) AS EdadIng,
Establecimientos.Tipo AS Est, Localidad.Loc
FROM MatAprAñoAcadCarr INNER JOIN (MatCursAñoAcadCarr INNER JOIN (UltimoAñoCarrCurs INNER JOIN (((MateriasCuatrimeAño INNER JOIN (Condicion INNER JOIN ((((((Establecimientos INNER JOIN
(Localidad INNER JOIN Alumnos ON Localidad.CodigoLocalidad = Alumnos.[Id Localidad Egre]) ON Establecimientos.Codigo = Alumnos.[Codigo Establecimiento]) INNER JOIN ((Carreras INNER JOIN
Matriculaciones ON Carreras.Departamento = Matriculaciones.Departamento) INNER JOIN [Carrera por Alumnos] ON Carreras.[Codigo Carrera] = [Carrera por Alumnos].[Codigo Carrera]) ON
Alumnos.Documento = [Carrera por Alumnos].Alumno) INNER JOIN MateriasCuatrimeAñoCondicion ON [Carrera por Alumnos].[Numero matricula] = MateriasCuatrimeAñoCondicion.Matricula) INNER
JOIN PromAño ON [Carrera por Alumnos].[Numero matricula] = PromAño.Matricula) INNER JOIN PromAprob ON [Carrera por Alumnos].[Numero matricula] = PromAprob.Matricula) INNER JOIN
MateriasCuatriC ON [Carrera por Alumnos].[Numero matricula] = MateriasCuatriC.Matricula) ON Condicion.CODIGO = [Carrera por Alumnos].Condicion) ON MateriasCuatrimeAño.Matricula = [Carrera
por Alumnos].[Numero matricula]) INNER JOIN Aprobacion ON [Carrera por Alumnos].[Numero matricula] = Aprobacion.Matricula) INNER JOIN MateriasRendAñoss ON [Carrera por Alumnos].[Numero
matricula] = MateriasRendAñoss.Matricula) ON UltimoAñoCarrCurs.Matricula = [Carrera por Alumnos].[Numero matricula]) ON MatCursAñoAcadCarr.Matricula = [Carrera por Alumnos].[Numero
matricula]) ON (MatAprAñoAcadCarr.Matricula = [Carrera por Alumnos].[Numero matricula]) AND (MatAprAñoAcadCarr.Matricula = [Carrera por Alumnos].[Numero matricula])
WHERE (((Carreras.Departamento) In ('ADMINISTRACION','INFORMÁTICA')) AND (([Carrera por Alumnos].[Numero matricula]) Between [Matriculaciones.Primera] And [Matriculaciones.Ultima]) AND
((Matriculaciones.[Año Ingreso]) Between 1999 And 2009) AND ((Matriculaciones.Sede) In (1)) AND ((Carreras.Años)>2))
GROUP BY Carreras.[Grupo Plan], Carreras.AreaDpto, Condicion.DESCRIPCION, UltimoAñoCarrCurs.MáxDeAño, MatCursAñoAcadCarr.[1], MatAprAñoAcadCarr.[1], MatCursAñoAcadCarr.[2],
MatAprAñoAcadCarr.[2], MateriasCuatrimeAño.[1999], MateriasCuatrimeAñoCondicion.[1999], MateriasRendAñoss.[1999], MateriasCuatrimeAño.[2000], MateriasCuatrimeAñoCondicion.[2000],
MateriasRendAñoss.[2000], MateriasCuatrimeAño.[2001], MateriasCuatrimeAñoCondicion.[2001], MateriasRendAñoss.[2001], MateriasCuatrimeAño.[2002], MateriasCuatrimeAñoCondicion.[2002],
MateriasRendAñoss.[2002], MateriasCuatrimeAño.[2003], MateriasCuatrimeAñoCondicion.[2003], MateriasRendAñoss.[2003], MateriasCuatrimeAño.[2004], MateriasCuatrimeAñoCondicion.[2004],
MateriasRendAñoss.[2004], MateriasCuatrimeAño.[2005], MateriasCuatrimeAñoCondicion.[2005], MateriasRendAñoss.[2005], MateriasCuatrimeAño.[2006], MateriasCuatrimeAñoCondicion.[2006],
MateriasRendAñoss.[2006], MateriasCuatrimeAño.[2007], MateriasCuatrimeAñoCondicion.[2007], MateriasRendAñoss.[2007], MateriasCuatrimeAño.[2008], MateriasCuatrimeAñoCondicion.[2008],
MateriasRendAñoss.[2008], MateriasCuatrimeAño.[2009], MateriasCuatrimeAñoCondicion.[2009], MateriasRendAñoss.[2009], MateriasCuatrimeAño.[2010], MateriasCuatrimeAñoCondicion.[2010],
MateriasRendAñoss.[2010], PromAprob.[Total de Calificacion], PromAño.[Total de Calificacion], PromAprob.[1], PromAño.[1], PromAprob.[2], PromAño.[2], Establecimientos.Tipo, Localidad.Loc, [Carrera
por Alumnos].[Numero matricula], [Carrera por Alumnos].[Fecha Ingreso], Alumnos.[Fecha Nacimiento], [Carrera por Alumnos].[Numero matricula], ([Carrera por Alumnos].[Fecha Ingreso]);
```

2. Filtrado de Datos

a) Limpieza de Datos

1. **Datos Faltantes**
   - Ej: Campos con valores nulos.
2. **Datos Erróneos**
   - Ej: PromSecundario: valor cero (0).
3. **Datos No normalizados**
   - Ej: Localidades registradas como “Posadas-Misiones”, “POSADAS-MISIONES”, “POSADAS”, “POSADAS.MISIONES”, “Posadas”, “PDAS”, “PDAS-MISIONES”, “PDAS-MNES”.

b) Transformación de Datos

Valores Existentes	Valores Nuevos
BAJA DEFINITIVA	B_DEF
BAJA TEMPORAL	B_TEMP
EN CURSO	CURS
EGRESADO	EGRE

Valores Nuevos	Descripción
99a01	Periodo año de ingreso 1999 al 2001
02a03	Periodo año de ingreso 2002 al 2003
04a05	Periodo año de ingreso 2004 al 2005
06a07	Periodo año de ingreso 2006 al 2007
08a09	Periodo año de ingreso 2008 al 2009

Valores Existentes	Valores Nuevos
S (Soltero)	S
C (Casado), V (Viudo), y D (Divorciado/a)	C

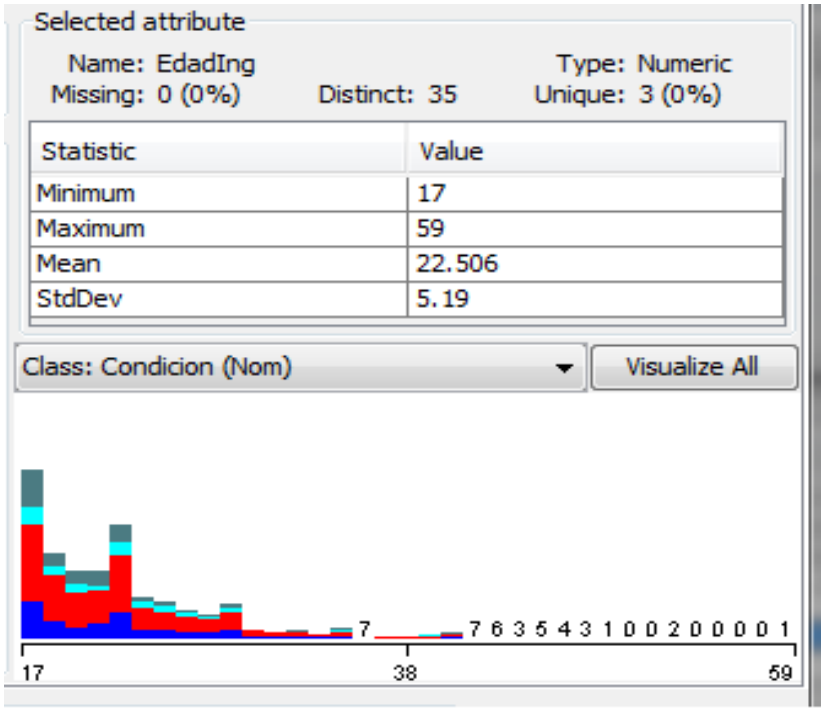
c) Estandarización e Integración de Datos

Transformaciones sintácticas

- Departamento: Dpto
- Grupo Carrera: GrupoC
- Carrera: Carr
- Año Ingreso: AñoIngr
- Edad Ingreso: EdadIng
- Sexo: Sexo
- Estado Civil: EstCiv
- Vive Padres: VcPadr
- Trabaja: Trab
- Establecimiento: Est
- Localidad: Loc
- Provincia: Prov
- Promedio General PromGralC
- Condición: Condición

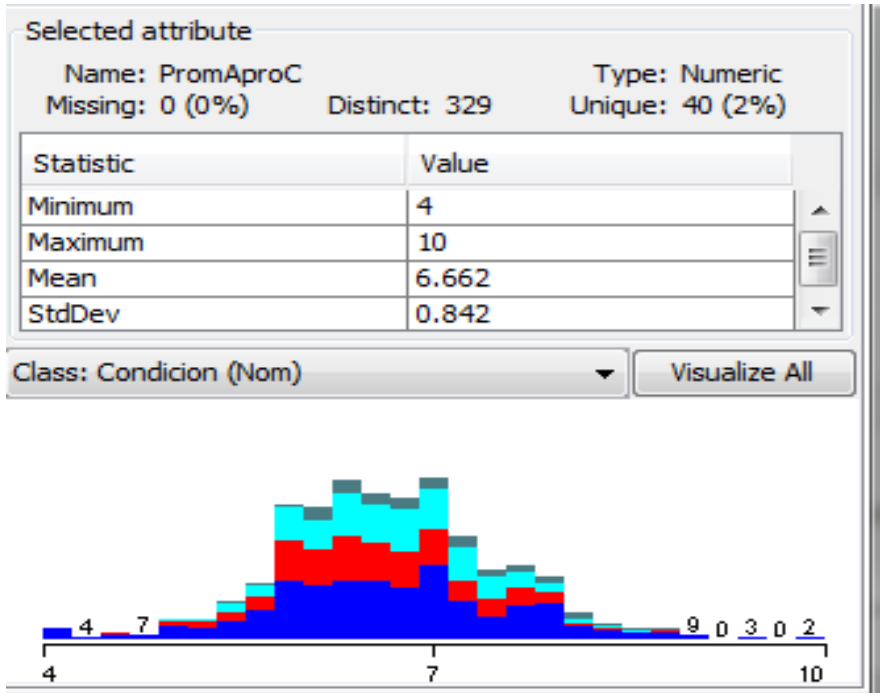
d) Análisis Explorativo de Datos Finales - Visualización

- Análisis de las distribuciones de valores, con los que se modifican o descartan algunos.



Nuevos Valores:

[17-19]
[20-22]
[23o+)



Nuevos Valores:

(4-6]
(6-7]
(7-10]

d) 2. Selección de Datos a Minar

Vistas Minables Finales en formato ARFF:

```
@attribute Plan {LCI98,CPN98,LAD98,LCO98,CPN05,LAD05,LTH05,LCI04,LAGH99,LTH01,LCO04,LGA04}
@attribute AñoAcad {1°a3°,4°a5°}
@attribute 1°Reg {[1-3],[4-5],[6-7],[8-9],[10o+)}
@attribute 1°Apr {[1-3],[4-7],[8-9],[10o+)}
@attribute 2°Reg {[1-3],[4-5],[6-7],[8-9],[10o+)}
@attribute 2°Apr {[1-3],[4-7],[8-9],[10o+)}
@attribute Curs1 {[1-5],[6-7],[8-9],[10-inf)}
@attribute FracC1 {[0],[1],[2],[3],[4-inf)}
@attribute Apro1 {[0-1],[2-3],[4-5],[6-inf)}
@attribute Curs2 {[1-5],[6-7],[8-9],[10-inf)}
@attribute FracC2 {[0],[1],[2],[3],[4-inf)}
@attribute Apro2 {[0-1],[2-3],[4-5],[6-inf)}
@attribute PromAproC {(4-6],(6-7],(7-10]}
@attribute PromGralC {(1-4],(4-5],(5-6],(6-7],(7-10]}
@attribute PromAp1 {(4-6],(6-7],(7-10]}
@attribute PromGr1 {(1-4],(4-5],(5-6],(6-7],(7-10]}
@attribute PromAp2 {(4-6],(6-7],(7-10]}
@attribute PromGr2 {(1-4],(4-5],(5-6],(6-7],(7-10]}
@attribute EdadIng {(17-19],[20-22],[23o+]}
@attribute Loc {OTRAS,INT_PROV,POSADAS}
@attribute Condicion {B_TEMP,B_DEF,EGRE,CURS}

@data
```

```
@attribute Plan {IIN98,LSI98,IIN04,PIN05}
@attribute AñoAcad {1°a3°,4°a5°}
@attribute 1°Reg {[1-3],[4-5],[6-7],[8-9],[10o+)}
@attribute 1°Apr {[1-3],[4-7],[8-9],[10o+)}
@attribute 2°Reg {[1-3],[4-5],[6-7],[8-9],[10o+)}
@attribute 2°Apr {[1-3],[4-7],[8-9],[10o+)}
@attribute Curs1 {[1-5],[6-7],[8-9],[10-inf)}
@attribute FracC1 {[0],[1],[2],[3],[4-inf)}
@attribute Apro1 {[0-1],[2-3],[4-5],[6-inf)}
@attribute Curs2 {[1-5],[6-7],[8-9],[10-inf)}
@attribute FracC2 {[0],[1],[2],[3],[4-inf)}
@attribute Apro2 {[0-1],[2-3],[4-5],[6-inf)}
@attribute PromAproC {(4-6],(6-7],(7-10]}
@attribute PromGralC {(1-4],(4-5],(5-6],(6-7],(7-10]}
@attribute PromAp1 {(4-6],(6-7],(7-10]}
@attribute PromGr1 {(1-4],(4-5],(5-6],(6-7],(7-10]}
@attribute PromAp2 {(4-6],(6-7],(7-10]}
@attribute PromGr2 {(1-4],(4-5],(5-6],(6-7],(7-10]}
@attribute EdadIng {(17-19],[20-22],[23o+]}
@attribute Loc {OTRAS,INT_PROV,POSADAS}
@attribute Condicion {B_TEMP,B_DEF,EGRE,CURS}

@data
```

3. Minería de datos

Técnicas Descriptivas

- Asociación
  - **Apriori**
- Segmentación
  - **EM**
  - **Cobweb (NB)**
  - **K-means**

Técnicas Predictivas

- Selección de Atributos
- Clasificación
  - **J48**
  - **OneR**

Resultados Característicos

1° y 2° Año

	INFORMÁTICA	ADMINISTRACIÓN
Año Académico	1° a 3° año	4° o 5° año
Edad de Ingreso	17 -19 años	23 años o mayores
Asignaturas Cursadas	Mayor cantidad	Menor cantidad
A. Cursadas/Fracasos	Medio/ Alto	Muy Bajo
A. Cursadas/Aprobadas	Baja/Media	Media
Promedios	Constantes	Disminución en 2° año
Características Adicionales	• A. Aprobadas en 1° → Año Académico. • 74% Baja Temp. → 4° o 5° año.	• 4° o 5° año, nivel actividad baja → Condición Egresado, Baja Temp o Definitiva.

Resultados y Conclusiones

- *Análisis comparativo y segmentado:*
  - Características similares y diferenciales → Acciones
- *Rendimiento académico en 2 tramos*:
  - Reducir niveles de clasificación
  - Obtener resultados agrupados más concretos
  - Ajustar los datos a las estructuras curriculares
- *Más de un atributo-indicador:*
  - *Éxito o fracaso académico:*
    - Calificaciones finales
    - Grado de éxito o fracaso de finalización de la carrera.

Aplicables a la Toma de Decisiones

- Incidencia de Promedios y N° de asignaturas Aprobadas
  - Medidas especiales de contención y eficacia del estudiante
- ***Dpto. de Administración***:
  - Estrategias de apoyo diferenciadas por edades;
  - Analizar contenidos, prácticas y metodologías de los espacios curriculares compartidos;
- ***Dpto. de Informática***:
  - Acciones de retención y apoyo a alumnos de los 1ros año;
  - Estrategias de motivación (finalización de la carrera);

Agricultura

“Aplicación de técnicas de minería de datos a un repositorio de variables fitofenológicas de cultivos cítricos”

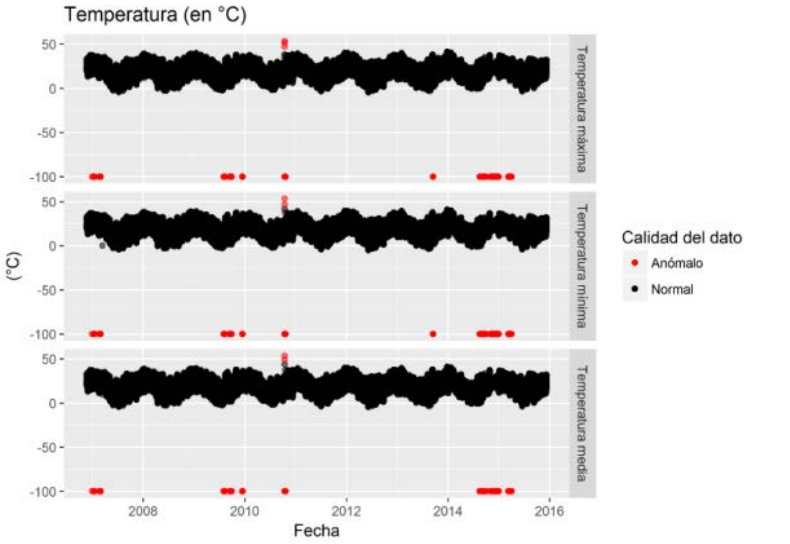
Objetivo

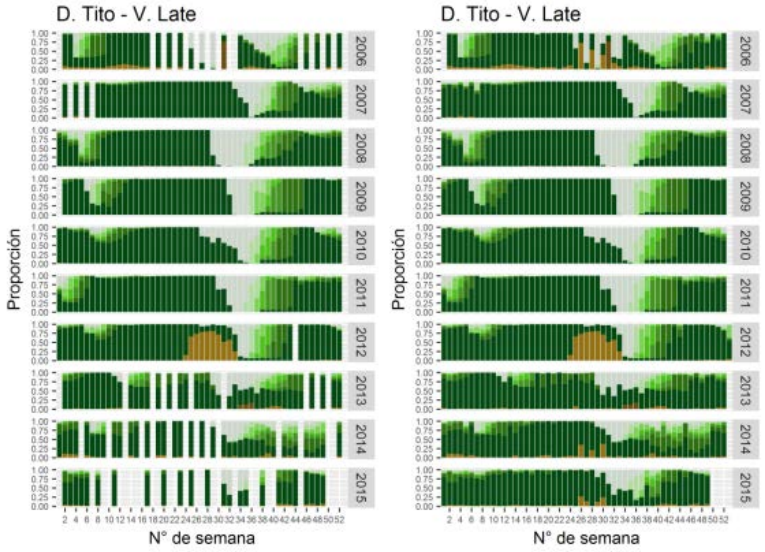
- Determinar las características que influyen en el desarrollo de los cultivos cítricos a través de las variables fitofenológicas y del triángulo de las enfermedades que son almacenadas en el sistema FruTIC, aplicando técnicas de minería de datos

  - Demás integrantes
    - Martín Ehman, Gabriel Surraco, Sergio Garran, Vanesa Hochmaier, Armando Taie

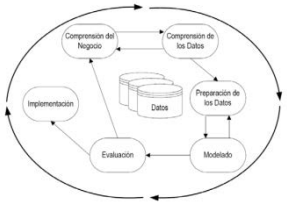
Agricultura

Atributo	Descripción
Estado	Estado general de la planta. Valores posibles: Malo, Regular, Bueno, Muy Bueno.
Brotación	Estadio de brotación de la planta observada. Valores posibles: B1, B2, B3, B3.4, B4, B5, B6, B7, B8.
Floración	Estadio de floración de la planta observada. Valores posibles: F0, F1.0, F1.1, F2, F3, F4, F5, F6, F7, F8.
Calibre	Diámetro ecuatorial del fruto. Valores mayores a cero.
MTD *Jackson* Mediterráneo	Cantidad de moscas del Mediterráneo por día en trampas *Jackson* Valores mayores o iguales a cero.
MTD *McPhail* Mediterráneo	Cantidad de moscas del Mediterráneo por día en trampas *McPhail* Valores mayores o iguales a cero.
MTD *McPhail* Americana	Cantidad de moscas americanas por día en trampas *McPhail* Valores mayores o iguales a cero.
Total minador	Cantidad total de ramas con presencia de minador en la planta observada.
Total *Diaphorina*	Cantidad total de ramas con presencia de *Diaphorina* en la planta observada.





Agricultura



Atributo	Métrica	Modelo ganador	Modelo seleccionado
Estado	Precisión	*random forest*	*xgboost*
	Kappa	*random forest*	
	AUC	*xgboost*	
Brotación	Precisión	*xgboost*	*xgboost*
	Kappa	*xgboost*	
	AUC	*xgboost*	
Floración	Precisión	C5.0	*xgboost*
	Kappa	C5.0	
	AUC	*xgboost*	
Calibre	RMSE	*random forest*	*random forest*
	R^2	*random forest*	
MTD *Jackson* Mediterráneo	RMSE	*xgboost*	*xgboost*
	R^2	*xgboost*	
MTD *McPhail* Mediterráneo	RMSE	*xgboost*	*xgboost*
	R^2	*xgboost*	
MTD *McPhail* Americana	RMSE	*xgboost*	*xgboost*
	R^2	*xgboost*	
Total minador	RMSE	*random forest*	*random forest*
	R^2	*random forest*	
Total *Diaphorina*	RMSE	*random forest*	*random forest*
	R^2	*random forest*	

Librerías externas

Paso de fronteras

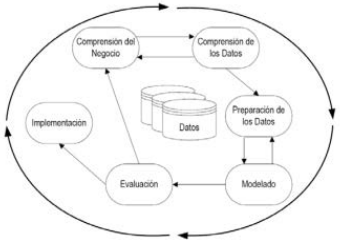
“Determinación de perfiles de clientes de un Centro Unificado de Frontera utilizando la combinación de Técnicas de Minería de Datos”

Objetivo

- Identificar cuáles son los perfiles de los clientes propensos a disminuir o aumentar la cantidad de cruces, en el Centro Unificado de Frontera (CUF) de Santo Tome Corrientes, a través de patrones de comportamiento obtenidos mediante la combinación de técnicas de Minería de Datos (MD)

  - Demás integrantes
    - Roque Ortega

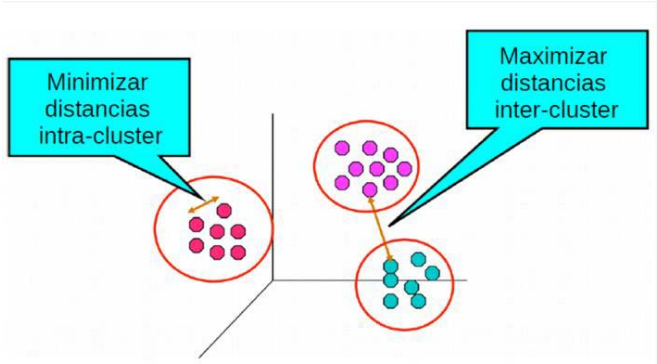
Paso de fronteras

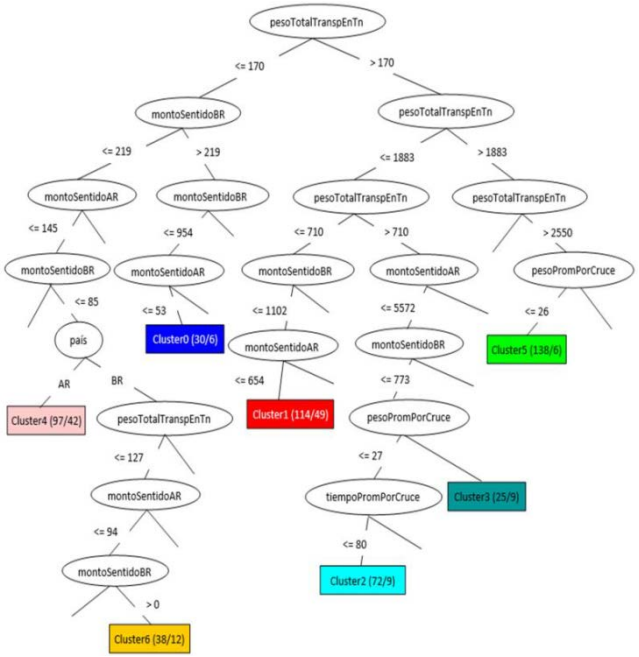


	R	F	M
5	Transacción más reciente	Más frecuente	Importe más alto
4			
3			
2			
1	Transacción menos reciente	Menos frecuente	Importe más bajo

Ejemplo: "5 4 1"

- Importe bajo
- Frecuencia de compra alta
- Transacción muy reciente





Detección de posibles fraudes

"Detección de pérdidas no técnicas en Sistemas de Distribución de Energía Eléctrica mediante herramientas de Data Science"

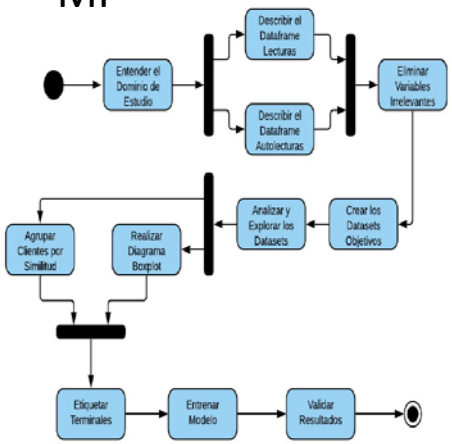
Objetivo

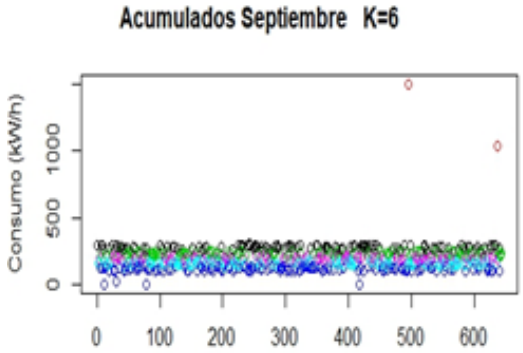
- Determinar los patrones de comportamiento de clientes con sospecha de fraude dentro del Sistema de Distribución de Energía Eléctrica (SDEE), mediante la implementación de técnicas de data science

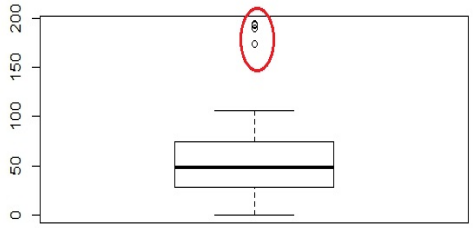
  - Demás integrantes
    - José Flores, Mariano Yavorski, Boris Da Silva

Detección de posibles fraudes

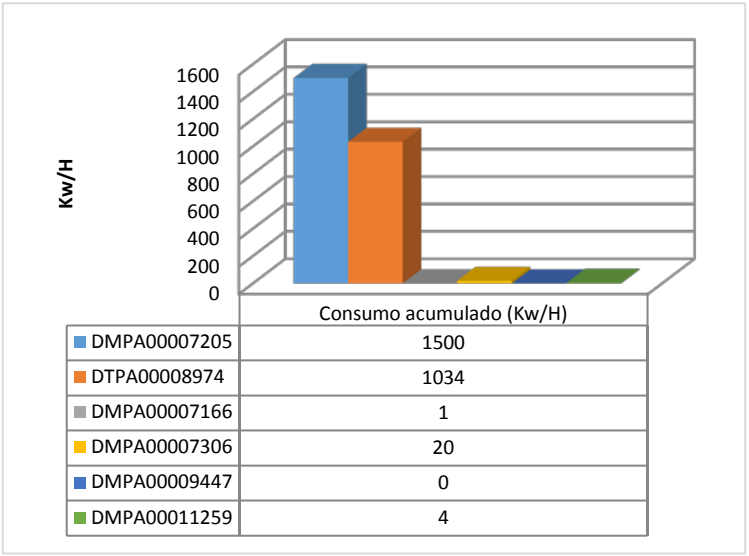
MP







Boxplot



XGBoost

Elecciones

“Análisis descriptivo de resultados electorales mediante técnicas de minería de datos”

Objetivo

- Describir mediante técnicas de minería de datos las características más representativas del electorado en relación a los resultados de las elecciones del año 2011 en la Provincia de Misiones”

  – Demás integrantes
    - John Neudeck, Alfredo Moyano

Minería Web, Semántica, PLN

- *"Modelo de Análisis de Información Desestructurada Utilizando Técnicas de Recopilación y Minería Web"* *(P-*
  - Demás integrantes
    - Marcelo Karanik, Roberto Suénaga, Fabián Favret, Tokuji Kairiyama , Victor Alvarenga, Matías Barboza, Leandro Witzke
- "Técnicas, herramientas y métodos semánticos para el procesamiento y recuperación de información documental jurídica" (NJS- SPARQL Stardog)
  - Demás integrantes
    - Héctor Ruidias, Juan Manuel Lezcano, Gabriel Dehner
- *"Clasificación de respuestas a consultas de disponibilidad hotelera a través de aprendizaje automático y procesamiento de lenguaje natural"* (P-*TensorFlow*--DL)
  - Demás integrantes
    - Emanuel Friedrich, Matías Koch

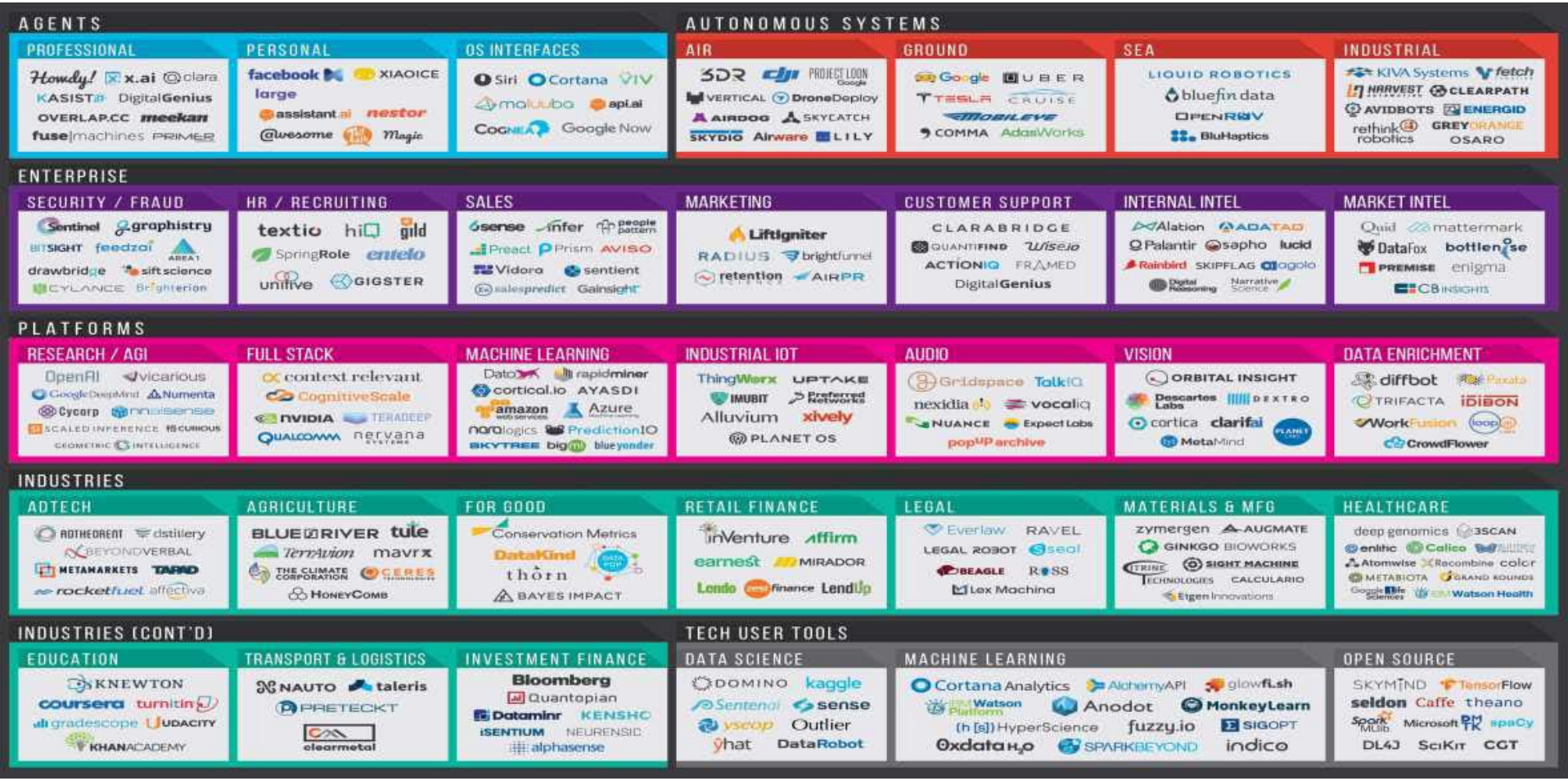
IN, PLN

- *“Inteligencia de Negocios aplicada a datos sobre Violencia Familiar”* (Pentaho)
  - Demás integrantes
    - Rodolfo Maggio, Nicolás Silvero, Germán Pouscht
- “*Extracción de Información de Historias Clínicas Digitales mediante Machine Learning*” (IBM Watson -ML)
  - Demás integrantes
    - Gabriel Candia, Sergio Montenegro , Nicolás López Forastier
- *“Análisis comparativo de técnicas de inteligencia artificial como soportes para la elaboración de diagnósticos clínicos”* (WEKA)
  - Demás integrantes
    - Mario Sotelo, Claudia Viera

Procesamiento de imágenes

- *"Clasificación de hojas de té al ingreso del proceso de secado a través de imágenes, mediante técnicas de inteligencia artificial"*
  - Demás integrantes
    - Luisina de Paula, Gabriel Guismín
- *"Reconocimiento de patrones de marcas en proyectiles de armas de fuego mediante procesamiento digital de imágenes y clasificación supervisada*"
  - Demás integrantes
    - Damián Dawidowicz , Manuel Quintana
- *"Detección de Senecio Brasiliensis al ingreso del té al secadero utilizando Máquina de Soporte Vectorial y procesamiento digital de imágenes"*
  - Demás integrantes
    - Federico Payes Alarcón, Fabián Favret

Herramientas y ecosistemas para Ciencia de Datos



Resumen

1

- **Algoritmos/técnicas**
  - Estructurados → ML: Sup. (predicción $) o No Sup. (descripción/segmentación)
  - Texto
  - Imágenes

2

- **Herramientas/Lenguajes**
  - Herramienta
  - Lenguaje: R o Python

3

- **Caso aplicación/uso**
  - contextualizar

¡MUCHAS GRACIAS!