El futuro digital es de todos

MinTIC

Blockchain & analítica de datos

para industrias digitales

etraining
Diseñamos nuevas formas de aprender

tecnalia Colombia
Inspiring Business

Analítica, BI, ML y AI

Presentación

Harry C Torres

Lic en matemáticas
Universidad Pedagógica Nacional
Estadístico
Universidad Nacional de Colombia
MSc en ingeniería de información
Universidad de los Andes
MSc en tecnologías de información para el negocio
Universidad de los Andes <en curso>

Experiencia en el sector financiero construyendo modelos de analítica avanzada para tomar decisiones.

Actualmente lidero la implementación de casos de uso basados en analítica avanzada en **AB-InBev – Bavaria**, en la vicepresidencia de marketing

Linkedin: https://www.linkedin.com/in/hctorresm/

Antes de comenzar

Fuga de clientes

Imagen tomada de: https://blog.serenacapital.com/saas-companies-how-to-reduce-churn-related-to-covid19-in-saas-companies-41a6c9872766

¿Sabías que existen 4 tipos de analítica?

Tipo de analítica	Pregunta a resolver
Analítica descriptiva	¿Qué está pasando?
Analítica diagnóstico	¿Por qué está pasando?
Analítica Predictiva	¿Qué es lo más probable que pase?
Analítica prescriptiva	¿Qué necesito hacer?

Inteligencia de negocios <BI>

Tipo de analítica	Pregunta a resolver
Analítica descriptiva	¿Qué está pasando?
Analítica diagnóstico	¿Por qué está pasando?
Analítica Predictiva	¿Qué es lo más probable que pase?
Analítica prescriptiva	¿Qué necesito hacer?

Inteligencia de negocios <BI>

Ejemplo problema fuga de clientes en una organización

Año mes	Clientes antiguos	Nuevos clientes	Clientes fugados	Stock	Crecimiento stock
1	1.839.500	62.000	61.600	1.839.900	0,020%
2	1.777.850	62.500	62.050	1.840.350	0,024%
3	1.777.820	63.000	62.530	1.840.820	0,026%
4	1.777.200	64.100	63.620	1.841.300	0,026%
5	1.777.300	64.500	64.000	1.841.800	0,027%
6	1.776.825	65.100	64.975	1.841.925	0,007%
7	1.776.435	65.600	65.490	1.842.035	0,006%
8	1.776.140	66.000	65.895	1.842.140	0,006%
9	1.775.925	66.300	66.215	1.842.225	0,005%
10	1.775.855	66.450	66.370	1.842.305	0,004%
11	1.775.880	66.500	66.425	1.842.380	0,004%
12	1.775.440	67.000	66.940	1.842.440	0,003%

La inteligencia de negocios (BI)

La inteligencia de negocios (BI)

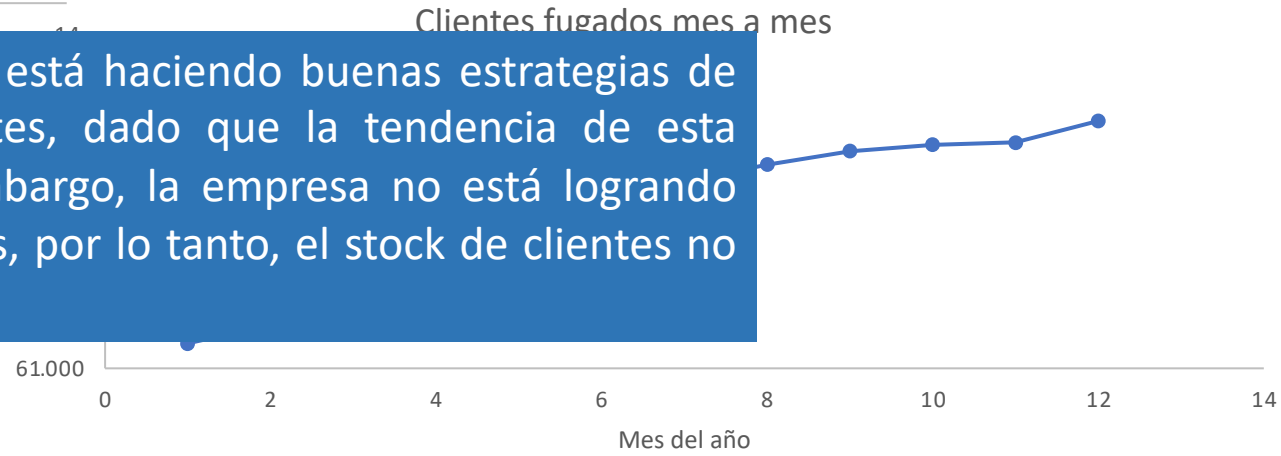
La inteligencia de negocios (BI)

etraining
Diseñamos nuevas formas de aprender

tecnalia Colombia
Inspiring Business

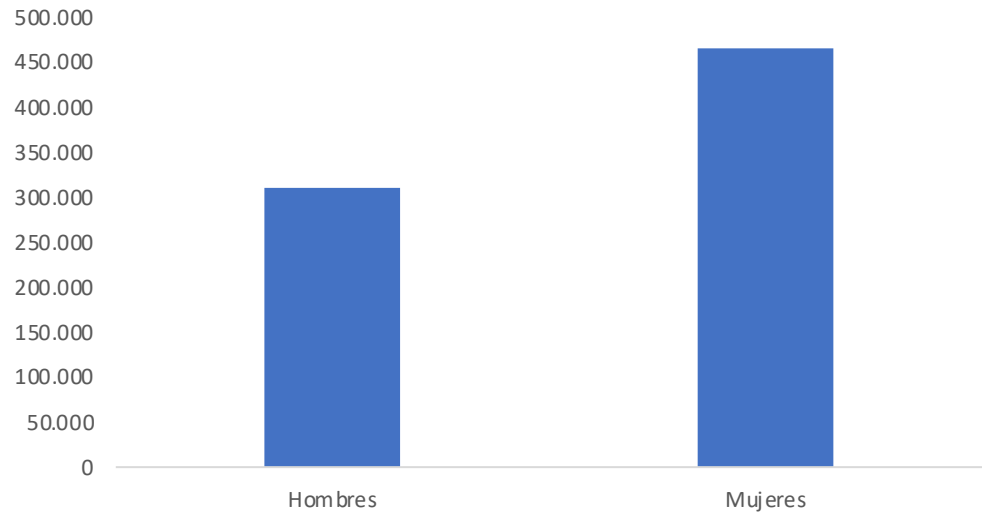
La inteligencia de negocios (BI)

Descubrimiento: La empresa está haciendo buenas estrategias de adquisición de nuevos clientes, dado que la tendencia de esta siempre es creciente, sin embargo, la empresa no está logrando retener a los clientes antiguos, por lo tanto, el stock de clientes no crece. **-> Analítica descriptiva**

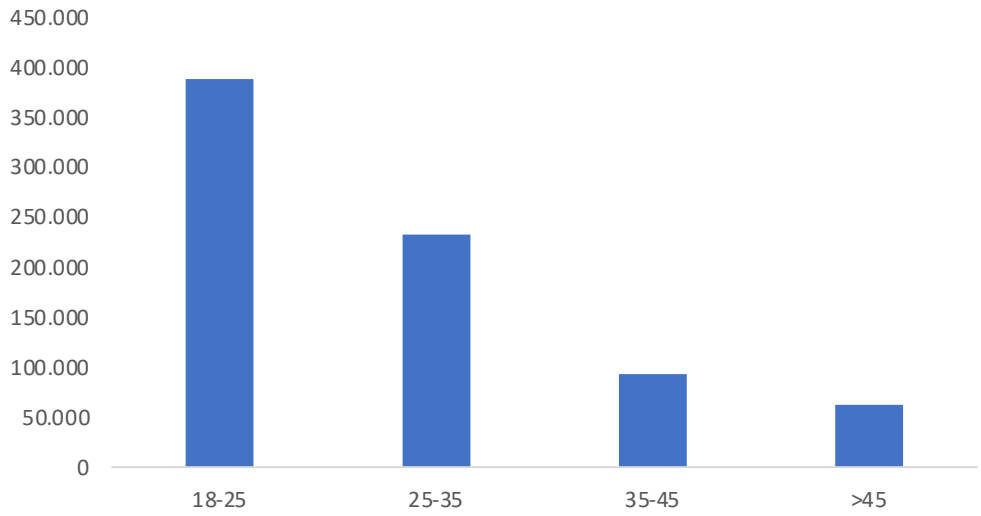


La inteligencia de negocios (BI)

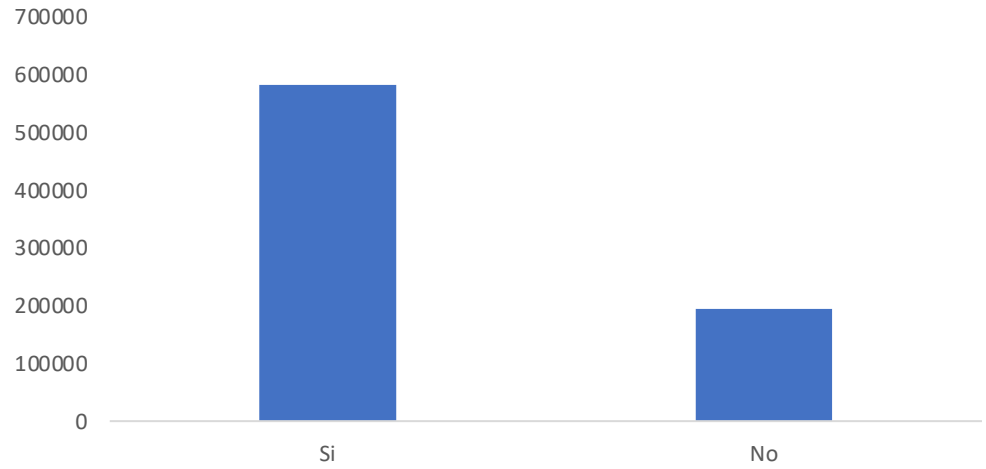
Género clientes que cancelan



Edad clientes que cancelan



Los clientes que se fugan presentan quejas o reclamos



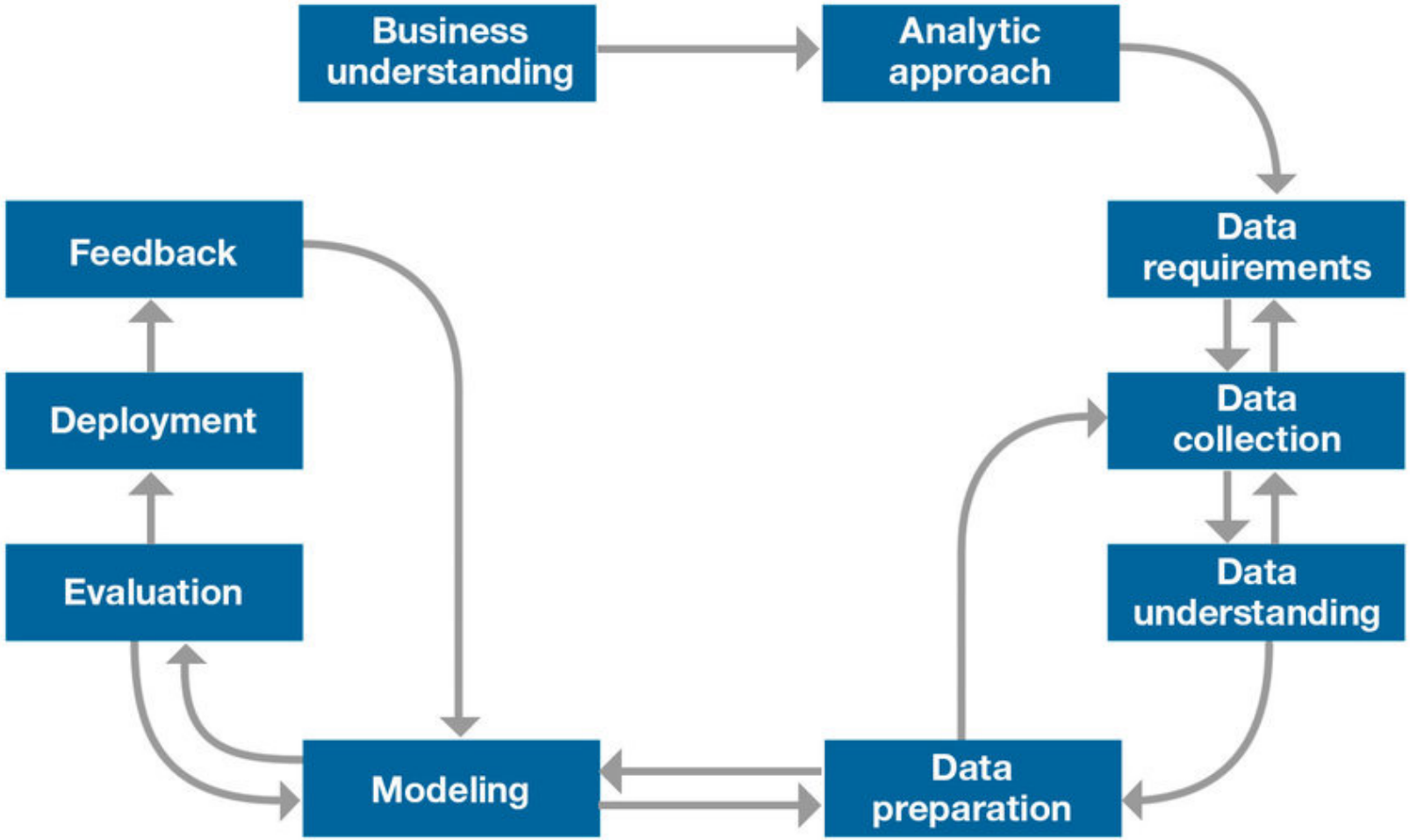
Descubrimiento: Las mujeres y los clientes jóvenes son aquellos que más han cancelado sus productos con la entidad.

- Muchos clientes han cancelado sus productos por que cuando estos presentaron una queja o un reclamo, no se solucionó oportunamente. **-> Analítica prescriptiva**

Machine Learning <ML>

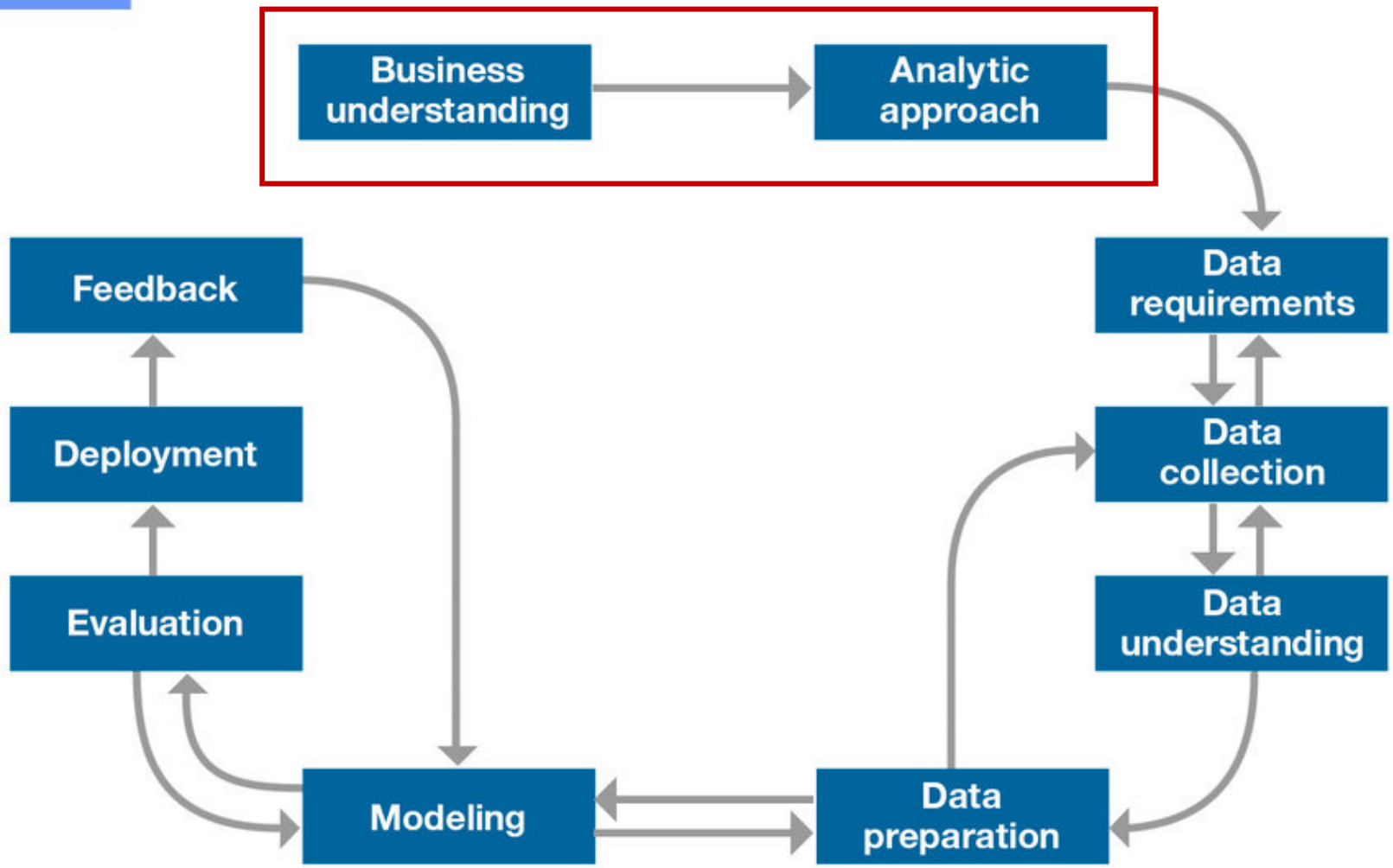
Tipo de analítica	Pregunta a resolver
Analítica descriptiva	¿Qué está pasando?
Analítica diagnóstico	¿Por qué está pasando?
Analítica Predictiva	¿Qué es lo más probable que pase?
Analítica prescriptiva	¿Qué necesito hacer?

Metodología Machine Learning <ML>



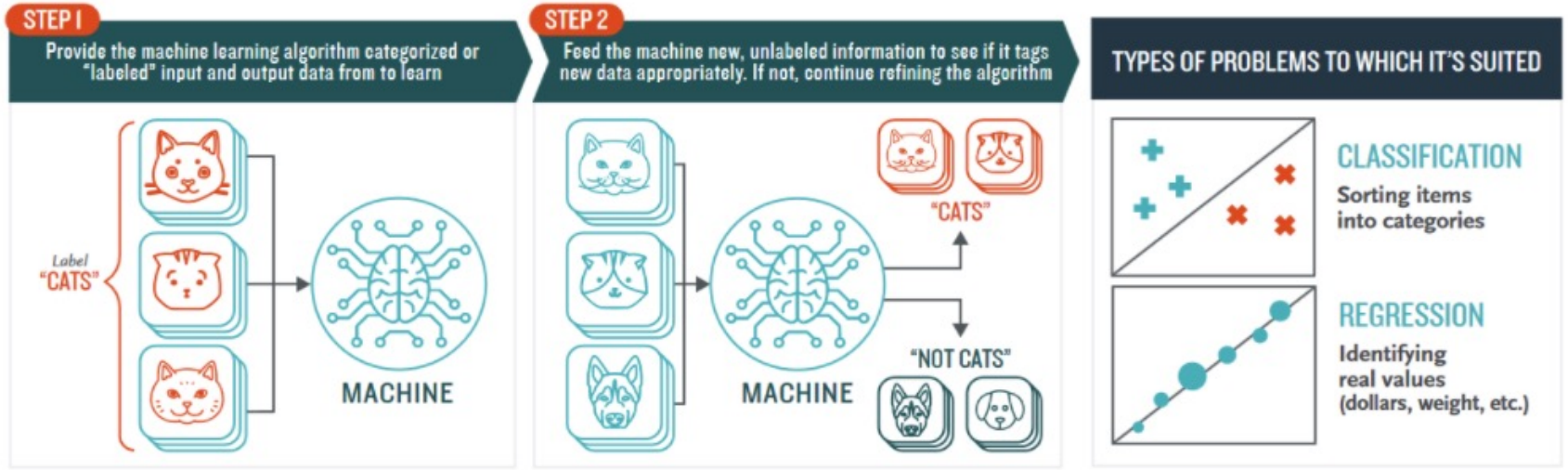
The IBM Foundational Methodology for Data Science. Source: [5].

Metodología Machine Learning <ML>

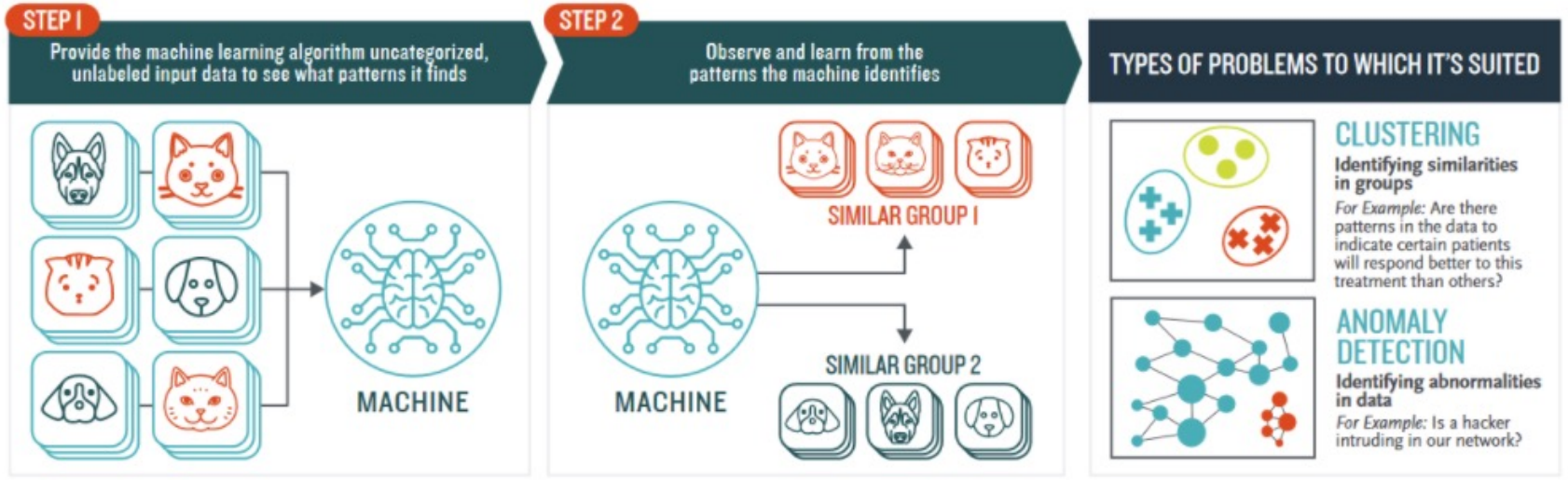


The IBM Foundational Methodology for Data Science. Source: [5].

Machine Learning <ML>



Modelos supervisados



Modelos no supervisados

El futuro digital es de todos | MinTIC

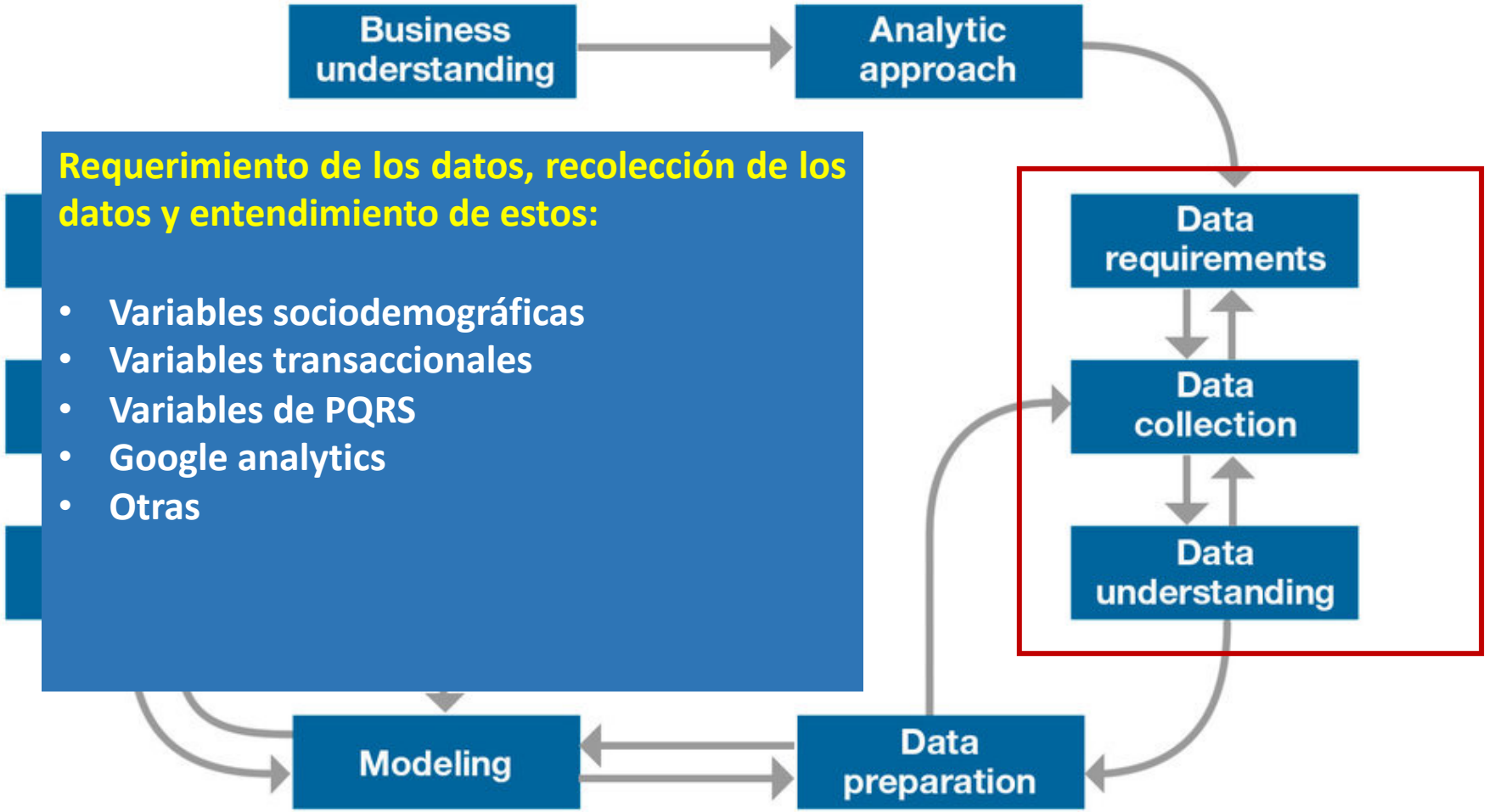
Machine Learning <ML>

etraining Diseñamos nuevas formas de aprender | tecnalia Colombia Inspiring Business

Metodología Machine Learning <ML>

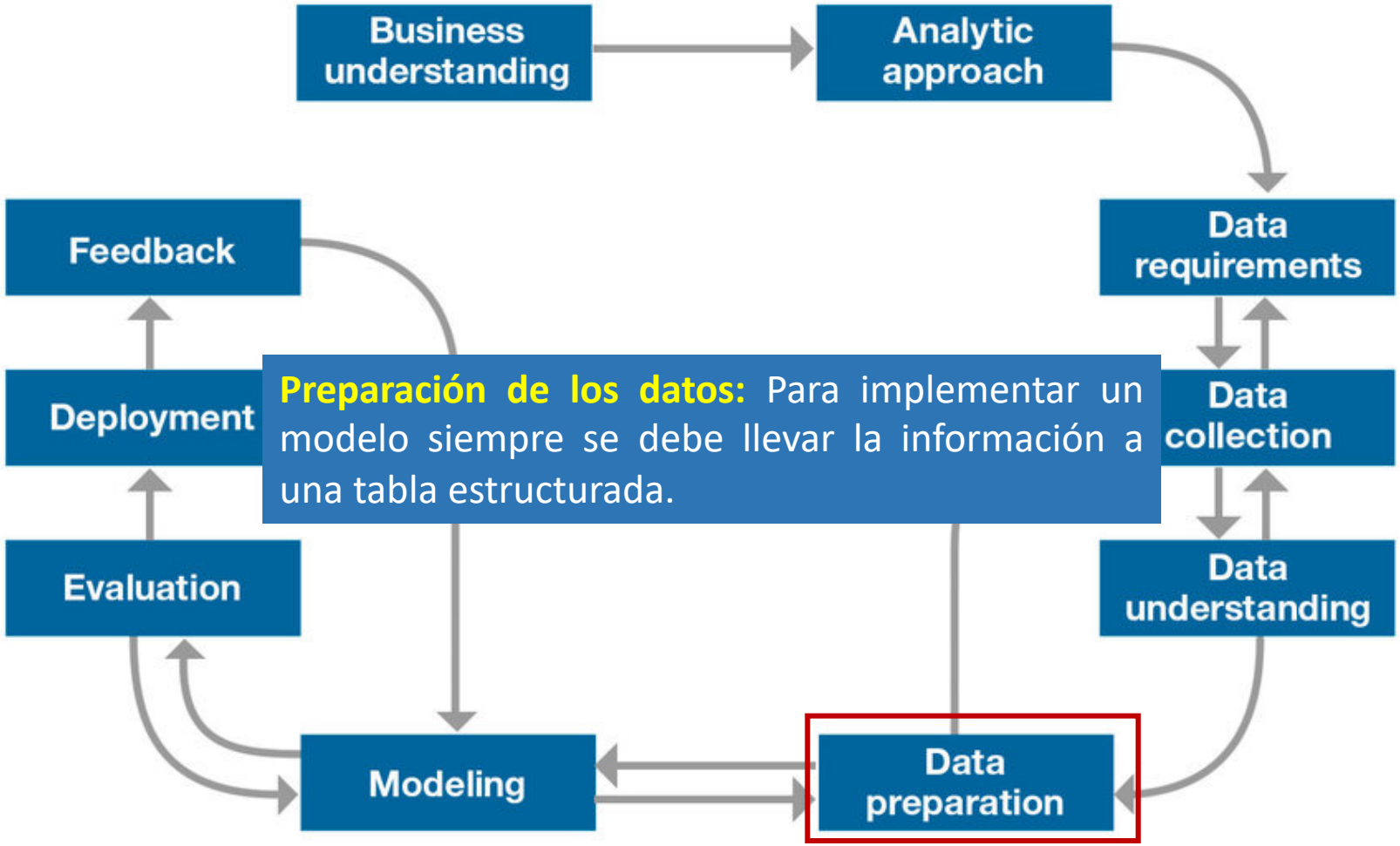
The IBM Foundational Methodology for Data Science. Source: [5].

Metodología Machine Learning <ML>



The IBM Foundational Methodology for Data Science. Source: [5].

Metodología Machine Learning <ML>



The IBM Foundational Methodology for Data Science. Source: [5].

ML: Preparación datos, fuga de clientes

Ejemplo 1 fuga de clientes: Se desea construir un modelo de aprendizaje supervisado para identificar los clientes más propensos a cancelar sus productos en los próximos **n** meses.

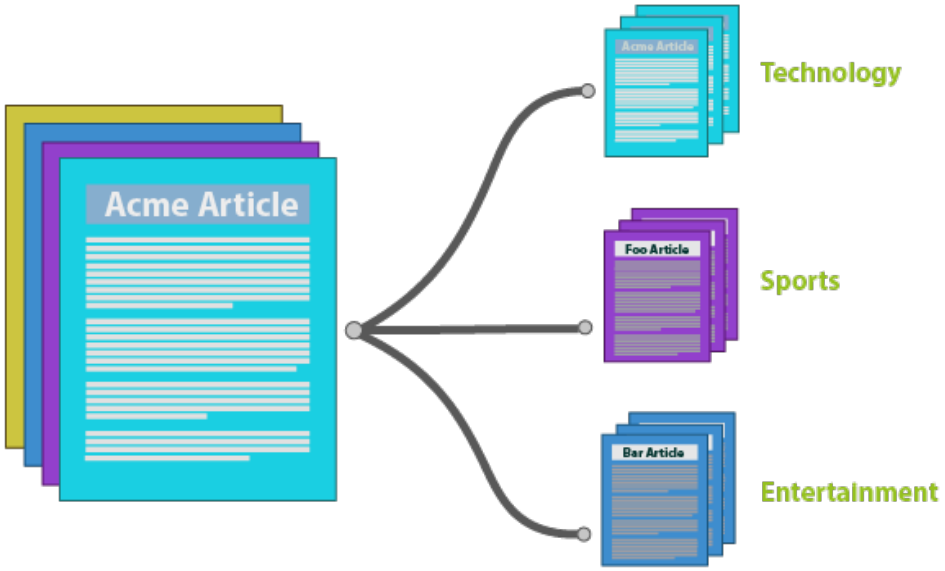
Cliente	Saldo TC	Saldo cta ahorros	…	Churn
1	1000000	5000000		0
2	0	100000		1

La información de los clientes debe corresponder a un mes determinado (ejemplo, Febrero 2019)

Se observa **n** meses después (Abril 2019) y se revisa si el cliente ha cancelado o no sus productos.

ML: Preparación datos, datos no estructurados

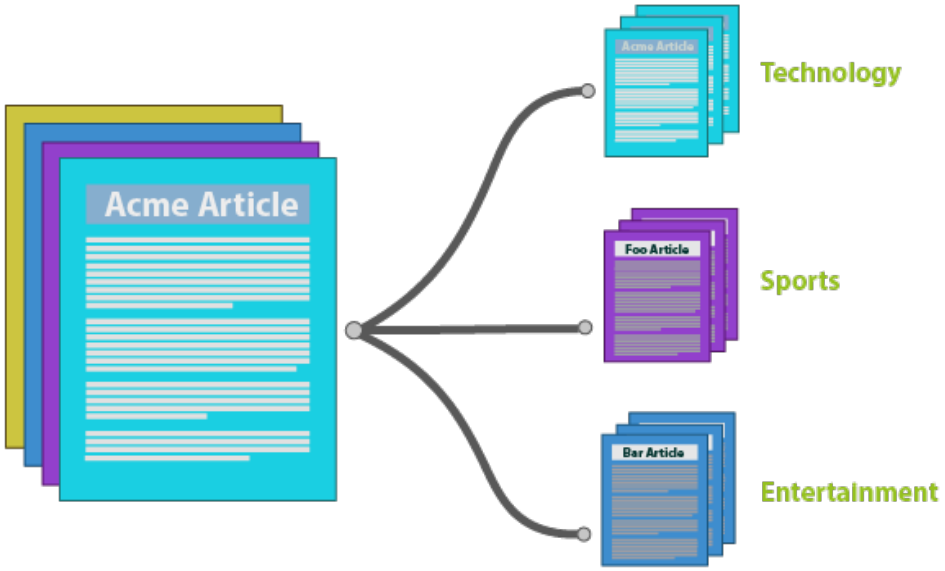
Ejemplo 2 clasificación de textos: Se desea construir un modelo de aprendizaje supervisado para clasificar documentos como tecnología, deportes o entretenimiento.



1. Convertir a minúsculas.
2. Eliminar caracteres especiales.
3. Eliminar stopwords
4. Lemmatizar

ML: Preparación datos, datos no estructurados

Ejemplo 2 clasificación de textos: Se desea construir un modelo de aprendizaje supervisado para clasificar documentos como tecnología, deportes o entretenimiento.



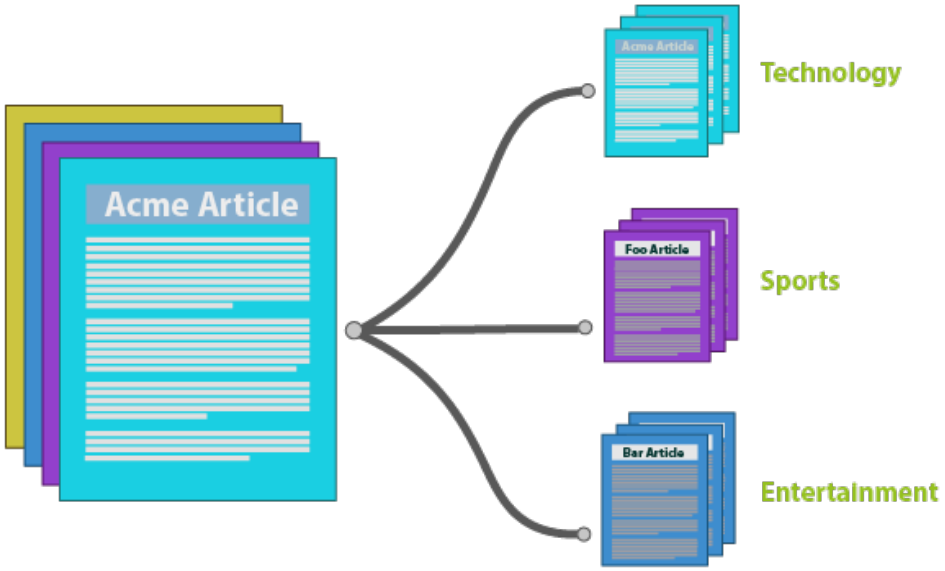
1. Convertir a minúsculas.
2. Eliminar caracteres especiales.
3. Eliminar stopwords
4. Lemmatizar

IDF =Log[(# Number of documents) / (Number of documents containing the word)] and

TF = (Number of repetitions of word in a document) / (# of words in a document)

ML: Preparación datos, datos no estructurados

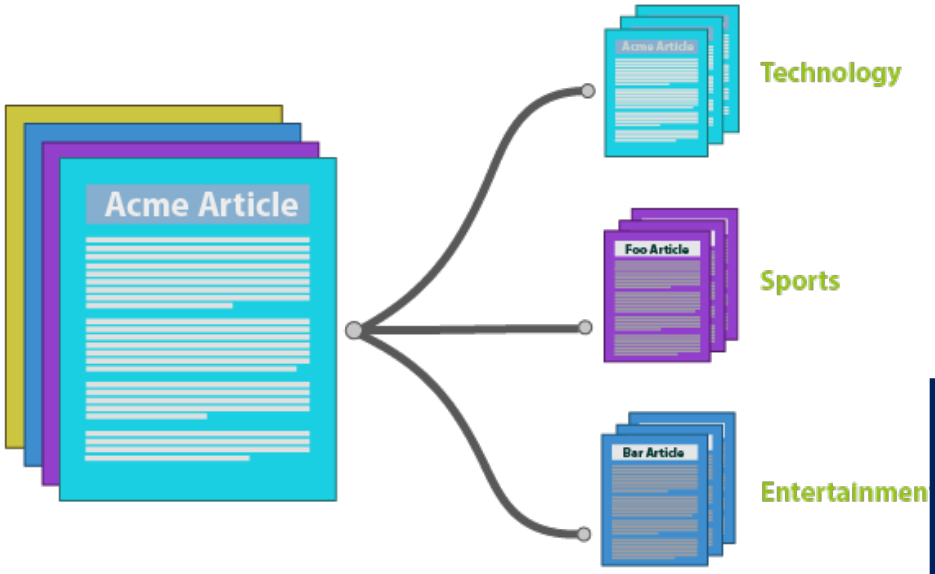
Ejemplo 2 clasificación de textos: Se desea construir un modelo de aprendizaje supervisado para clasificar documentos como tecnología, deportes o entretenimiento.



Documento	FIFA	go	rock	…	Tipo
1	0	0	0.2		0
2	0.4	0	0		1

ML: Preparación datos, datos no estructurados

Ejemplo 2 clasificación de textos: Se desea construir un modelo de aprendizaje supervisado para clasificar documentos como tecnología, deportes o entretenimiento.

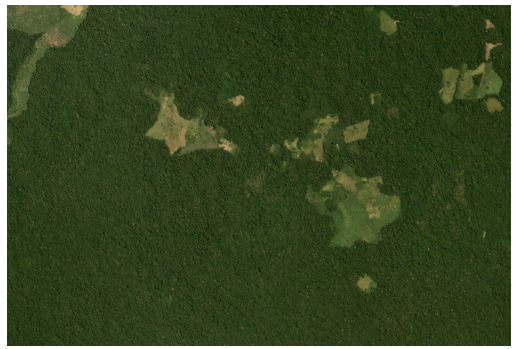


Documento	FIFA	go	rock	…	Tipo
1	o	0	0.2		0
2	0.4	0	0		1

En los datos, siempre se debe tener identificado la variable objetivo o target, tipo 0 = entretenimiento, tipo 1 = deportes, tipo 2 = tecnología

ML: Preparación datos, datos no estructurados

Ejemplo 3 clasificación de imágenes satelitales: A partir de dos imágenes satelitales tomadas a un mismo lugar y en diferentes momentos identificar si hay o no deforestación



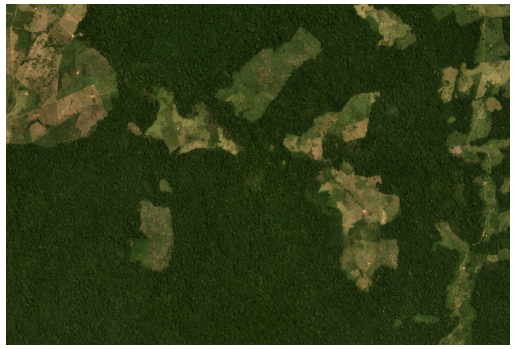
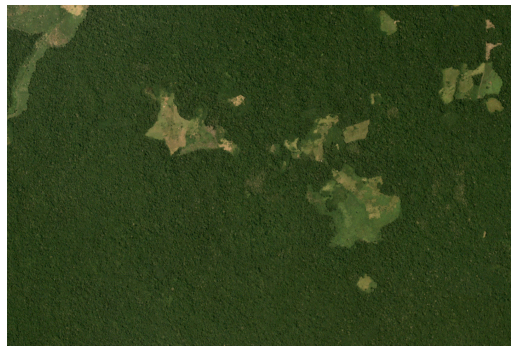


Imagen Parque nacional Chiribiquete, en el amazonas colombiano Agosto 2017 – Agosto 2019. Las imágenes presentan que hubo un proceso de deforestación.

A partir de distintas herramientas (ejemplo, OpenCV) se pueden extraer distintas características de las imágenes (ejemplo, diferencia de pixeles).

ML: Preparación datos, datos no estructurados

Ejemplo 3 clasificación de imágenes satelitales: A partir de dos imágenes satelitales tomadas a un mismo lugar y en diferentes momentos identificar si hay o no deforestación



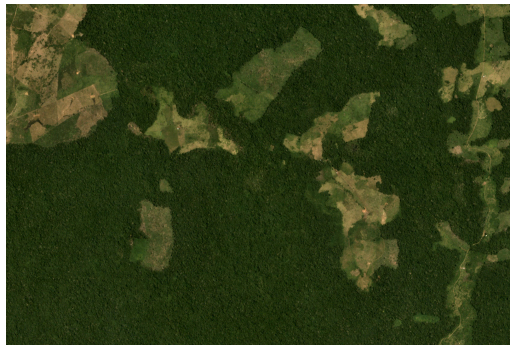
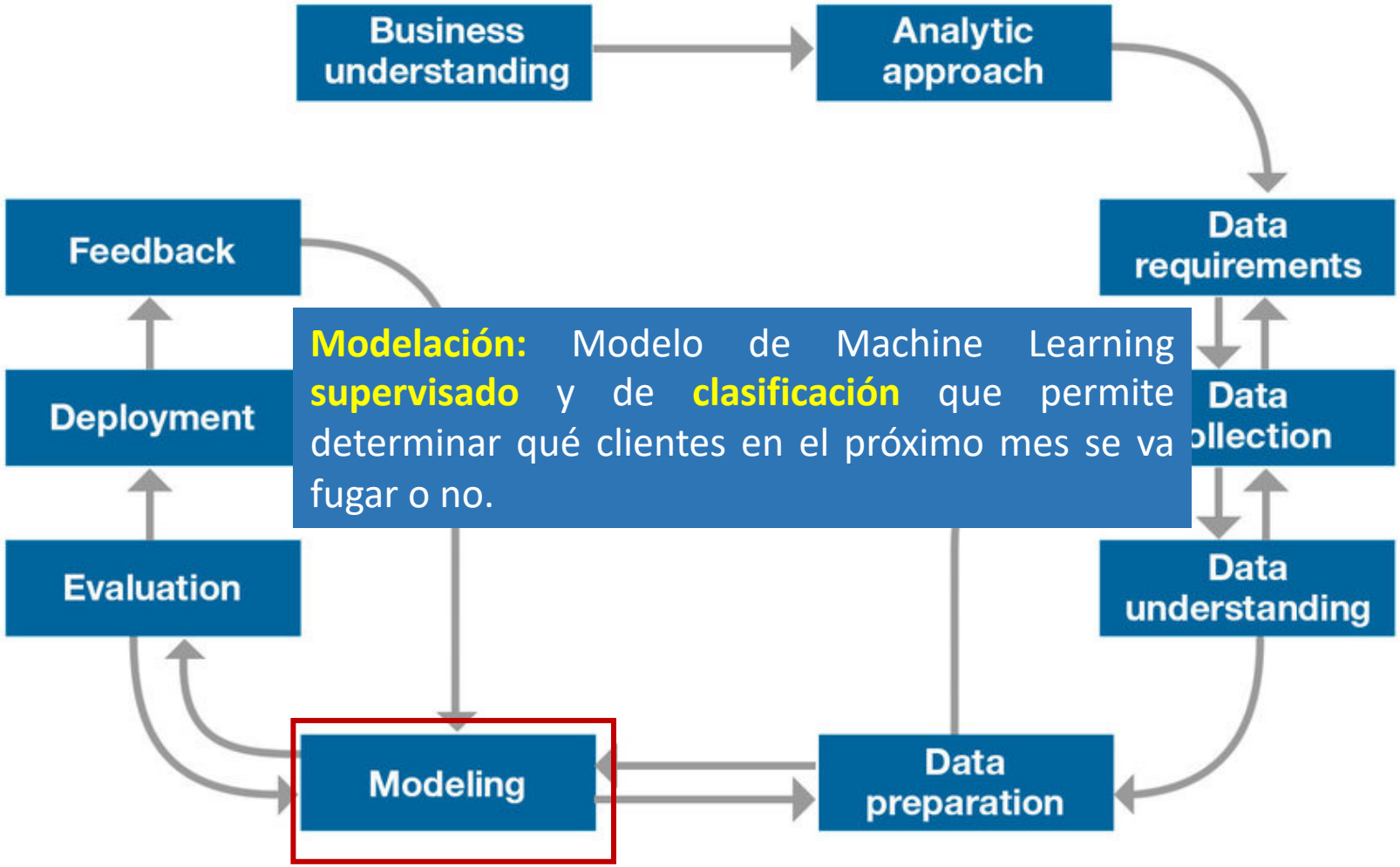


Imagen Parque nacional Chiribiquete, en el amazonas colombiano Agosto 2017 – Agosto 2019. Las imágenes presentan que hubo un proceso de deforestación.

Área	DP1	DP2	DP3	…	Tipo
1	0.2	0.7	0.9		1
2	0.002	0.003	0.004		0

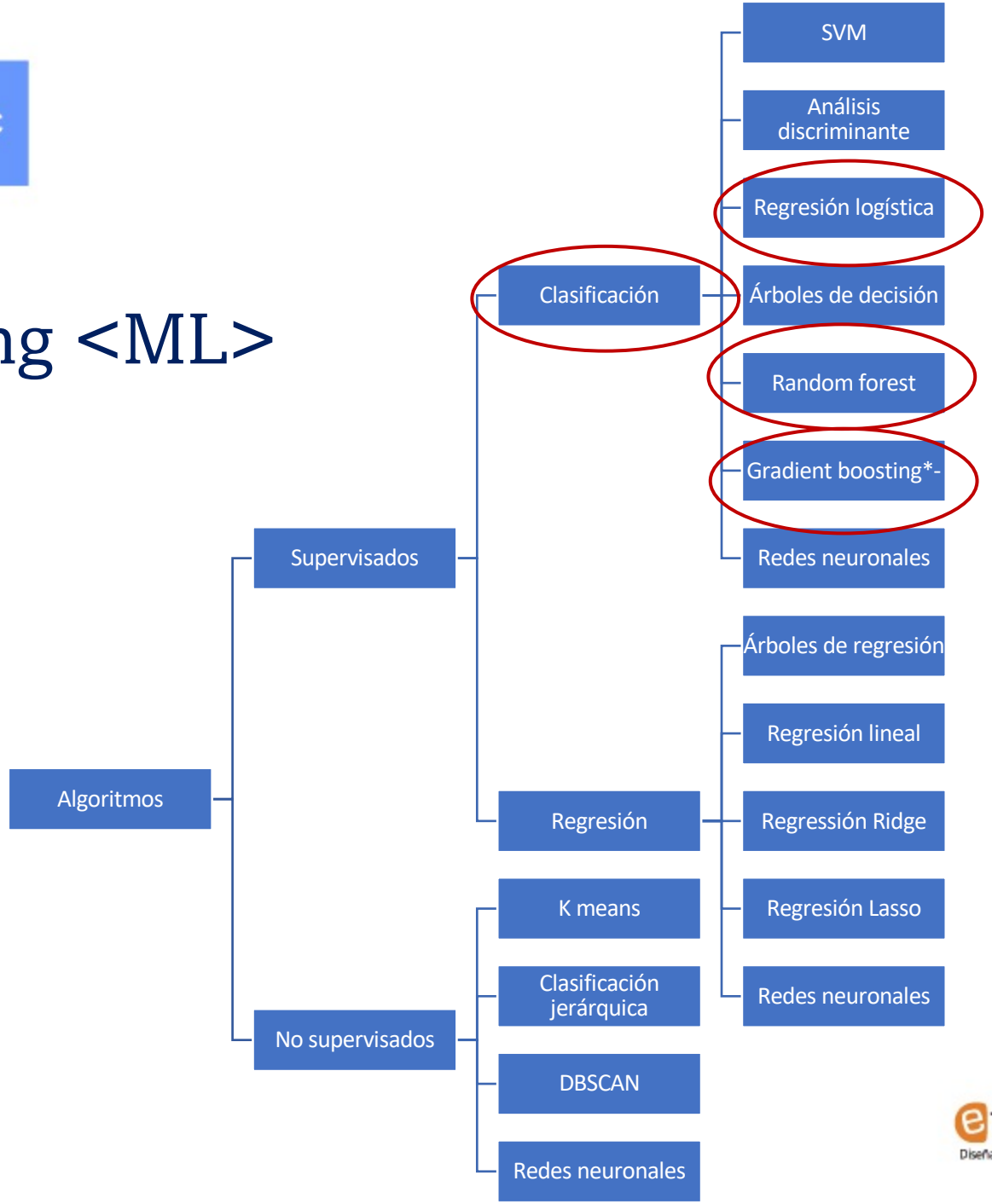
En los datos, siempre se debe tener identificado la variable objetivo o target, tipo 0 = si las imágenes no presentaron cambio, tipo 1 caso contrario.

ML - Fuga de clientes



The IBM Foundational Methodology for Data Science. Source: [5].

Machine Learning <ML>

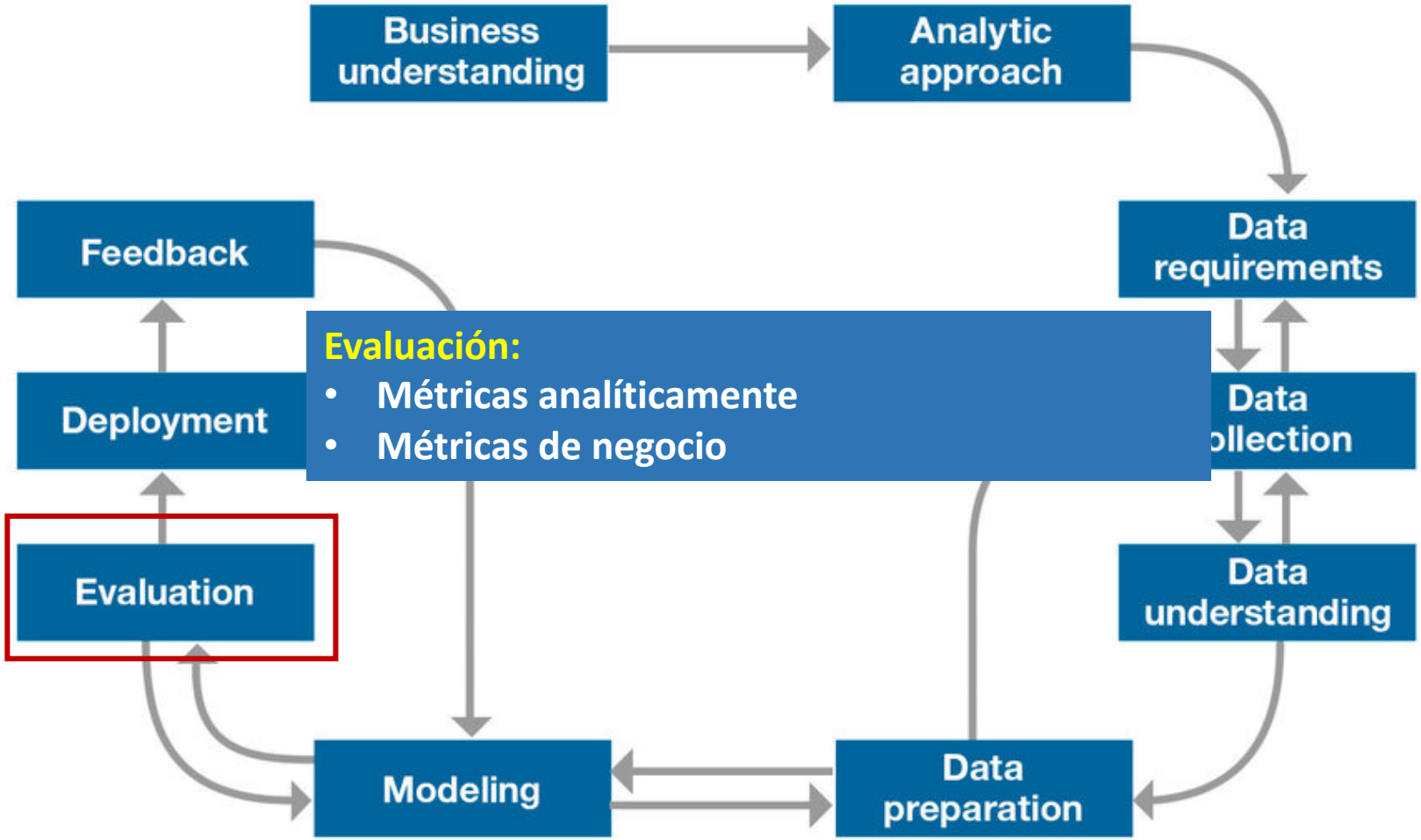


Interpretable

Robusto a datos atípicos y faltantes

Computacionalmente eficiente

Machine Learning <ML>



The IBM Foundational Methodology for Data Science. Source: [5].

ML - Evaluación

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Matriz de confusión:

Determinar el número total de individuos clasificados correcta e incorrectamente.

ML - Evaluación

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Accuracy:

Porcentaje de clientes que el modelo ha logrado clasificar correctamente

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Negative + False\ Positive + False\ Negative + True\ Positive}$$

ML - Evaluación

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Precision:

De todos los clientes que el modelo pronosticó que iban a cancelar sus productos, cuál es el porcentaje de estos que el modelo pronosticó correctamente.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

ML - Evaluación

		Predicted: Negative	Predicted: Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted: Negative	Predicted: Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Predicted: Negative	Predicted: Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

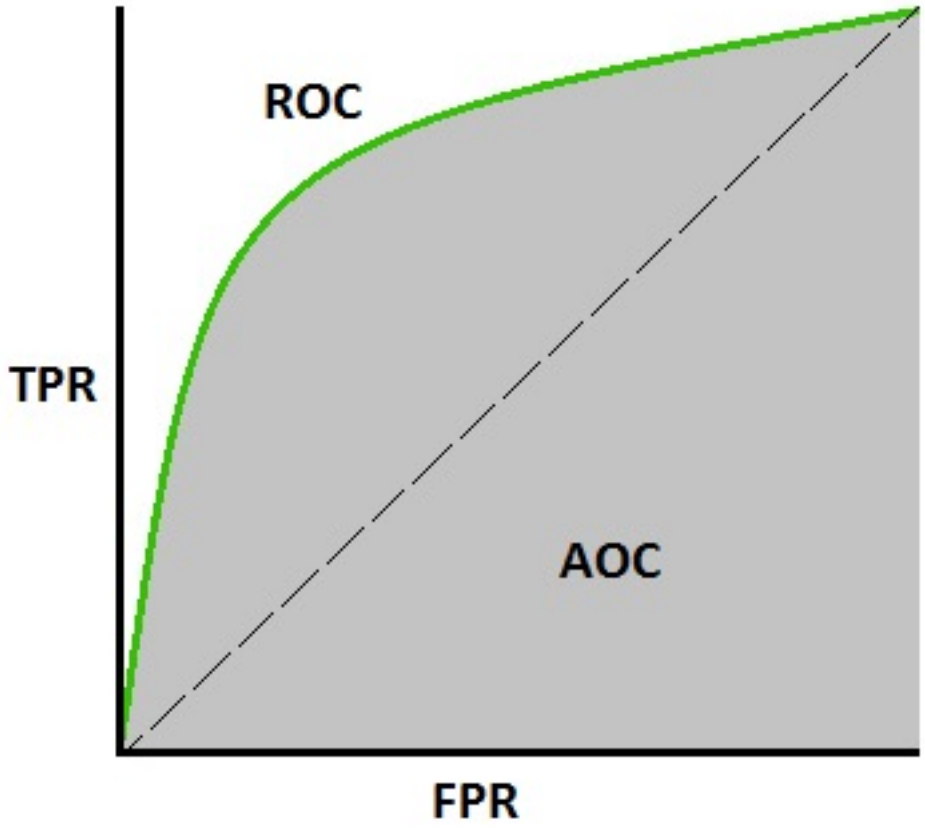
		Predicted: Negative	Predicted: Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Recall:

Del total de clientes que cancelaron sus productos, cuál es el porcentaje de estos que el modelo ha encontrado.

$$Recall = \frac{True\ Positive}{False\ Negativa + True\ Positive}$$

ML – Evaluación – ROC -AUC



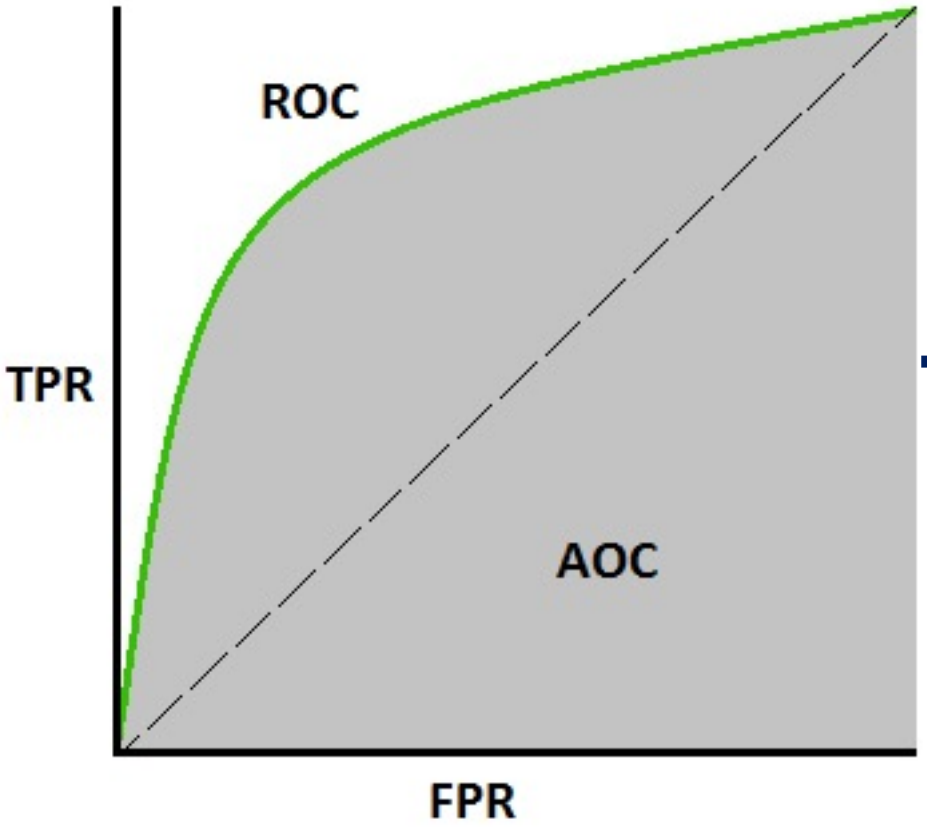
La curva ROC representa una relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) a diferentes puntos de corte.

$$TPR = Recall = \frac{True\ Positive}{False\ Negativa + True\ Positive}$$

$$FPR = 1 - \frac{True\ Negative}{False\ Positive + True\ Negative}$$

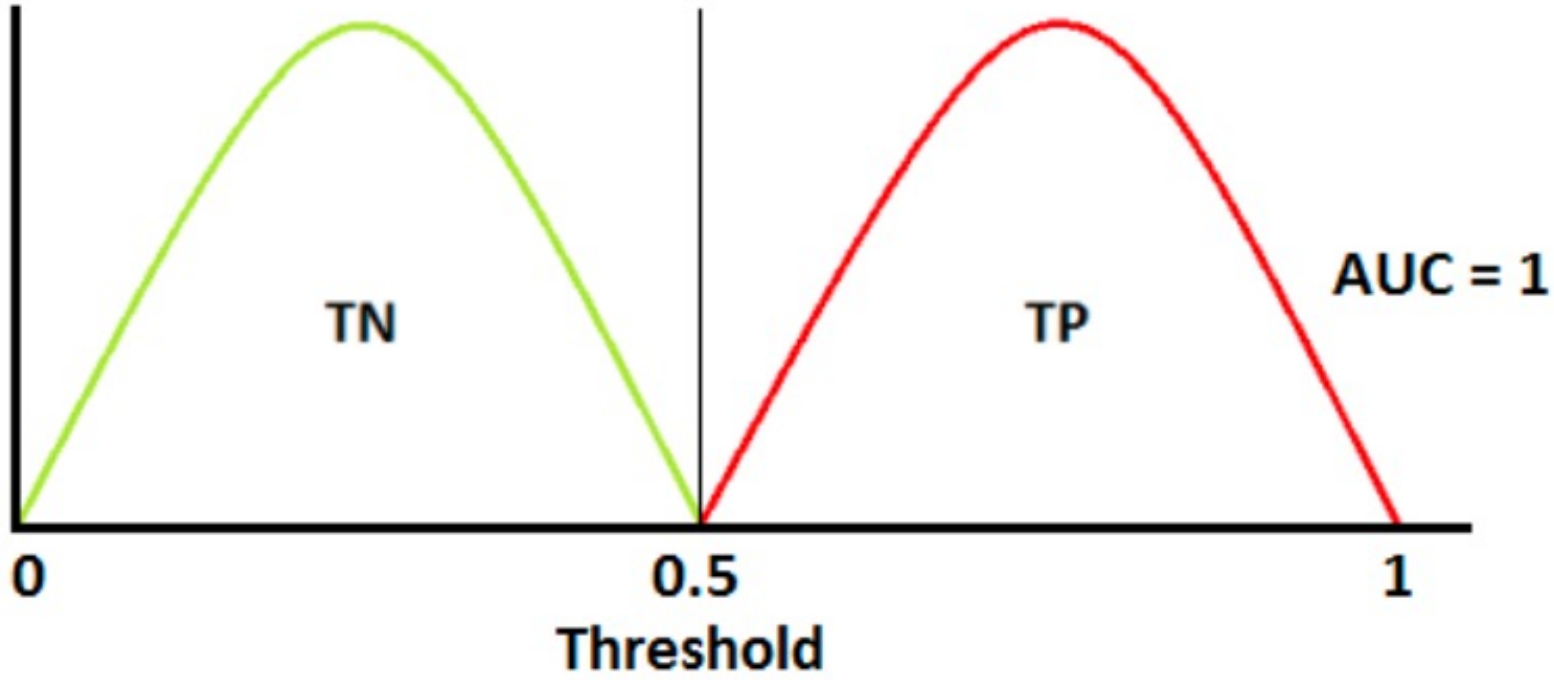
Note que el FPR es 1 – Recall (clase negativa)

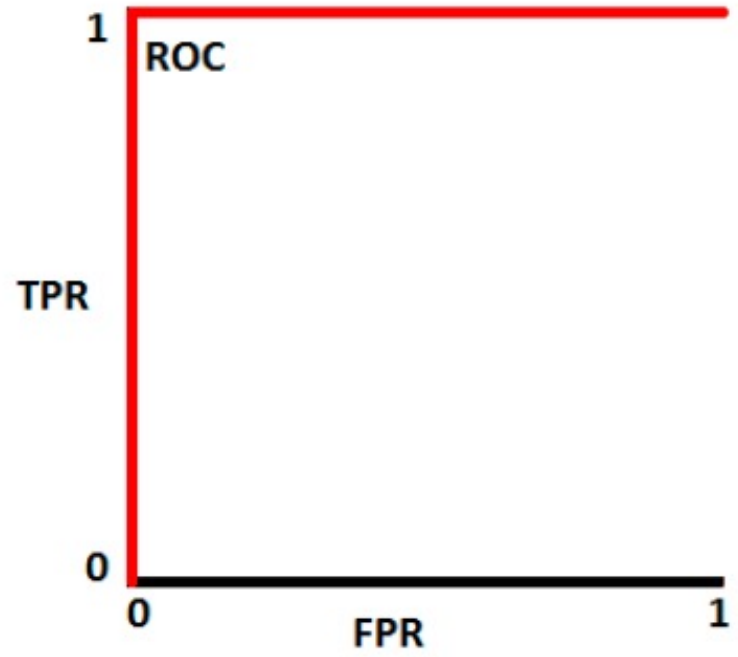
ML – Evaluación – ROC -AUC



Al calcular el área bajo la curva se obtiene el AUC, entre más grande sea este valor, mayor es la capacidad predictiva del modelo.

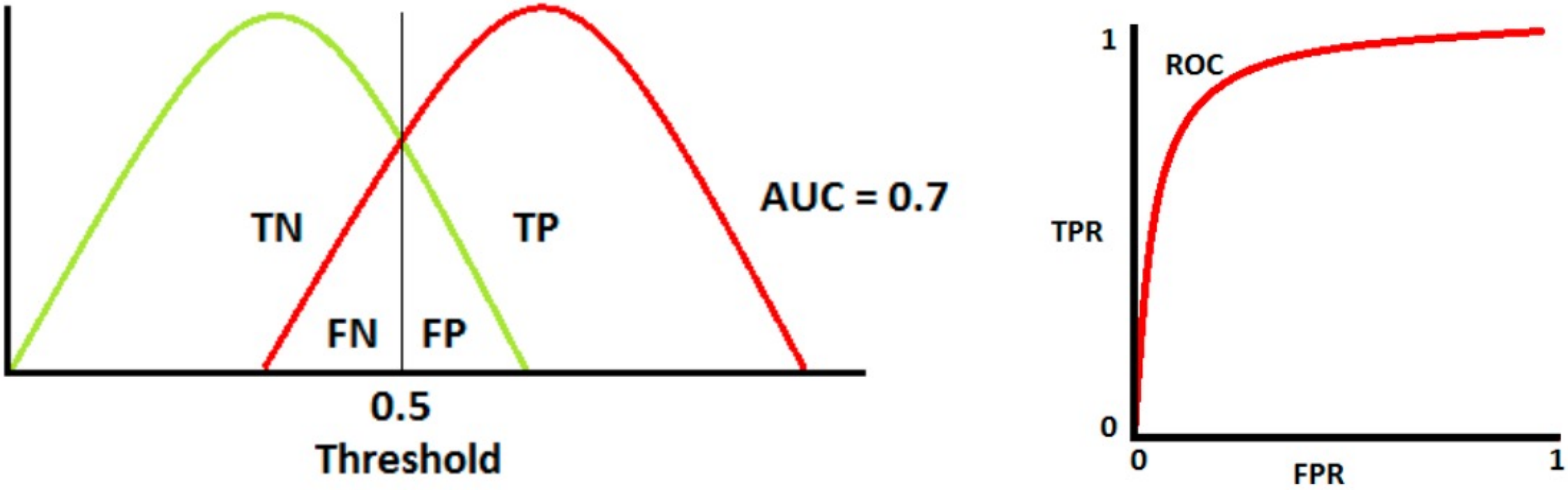
ML – Evaluación – ROC -AUC





El modelo es perfecto cuando AUC es igual a 1, si esto sucede, **sospeche!**

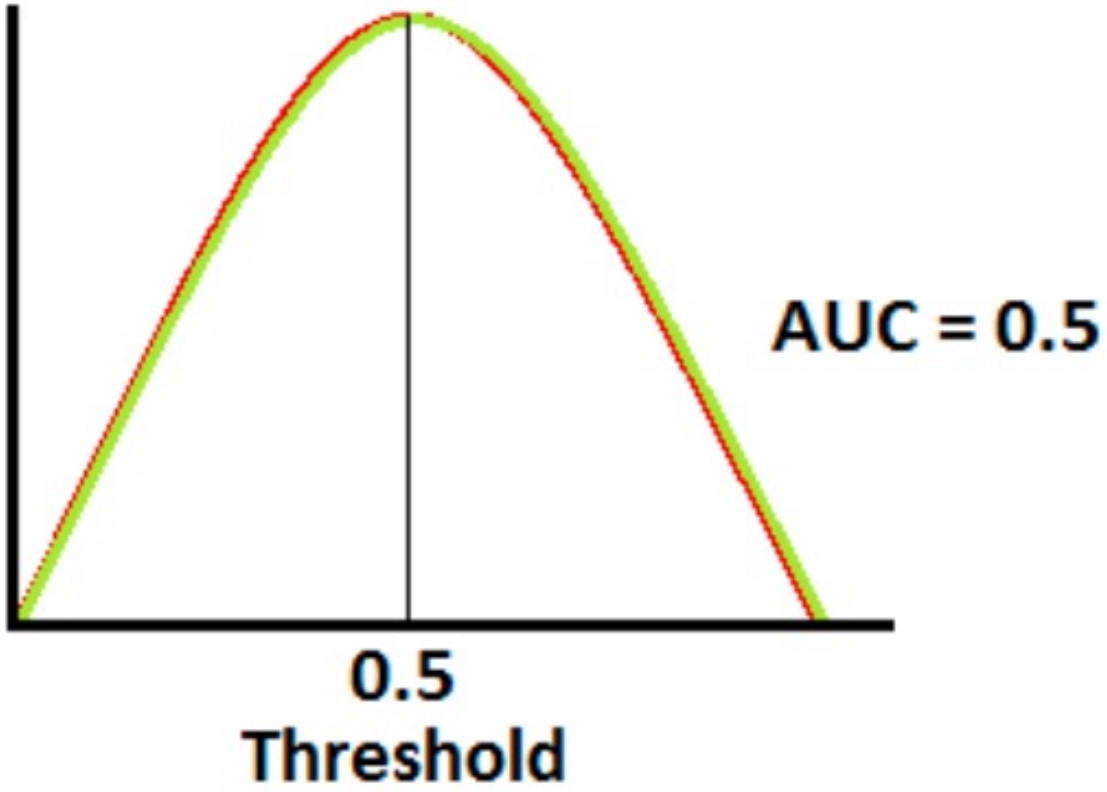
ML – Evaluación – ROC -AUC

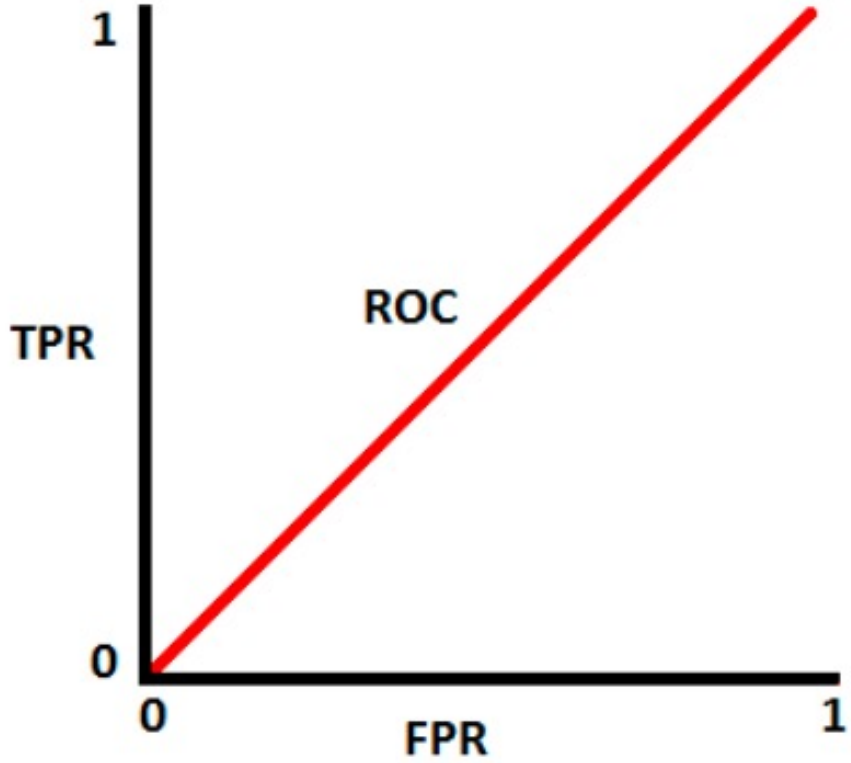


[Image 8 and 9] (Image courtesy: My Photoshopped Collection)

Un buen modelo presenta un AUC mayor o igual a 0.7.

ML – Evaluación – ROC -AUC





Un modelo no pronostico bien cuando el AUC es cercano a 0.5.

Los dos conceptos más importantes en ML

- En la estimación de un modelo el aspecto más importante a tener en cuenta es aprender las características esenciales de los datos y **no** una representación exacta de los datos.

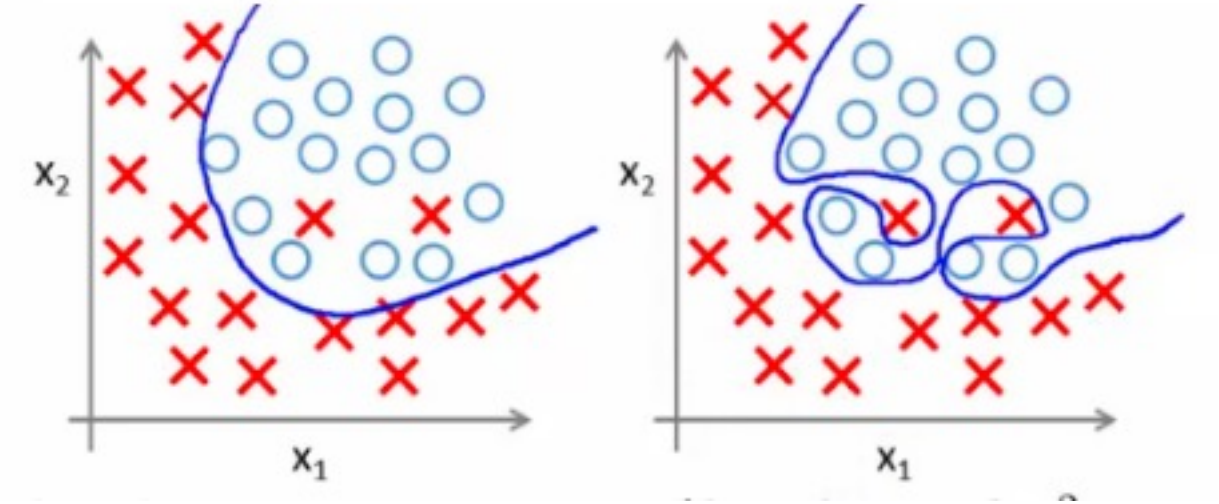
- Lo más importante en un modelo es garantizar la capacidad de generalización.

Sobreajuste <overfitting>

Esta casa se adecua perfectamente al tamaño del perro, sin embargo, esta no funciona con otros perros. Lo mismo sucede con los datos!

Sobreajuste <overfitting>

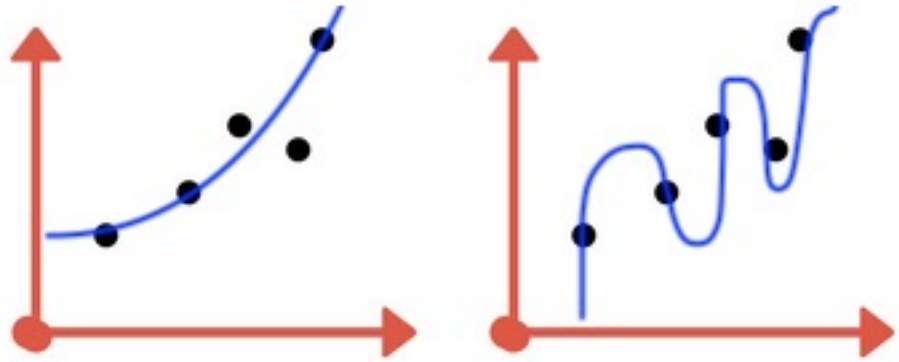
Modelo de clasificación



Buen ajuste ajuste

Sobre ajuste

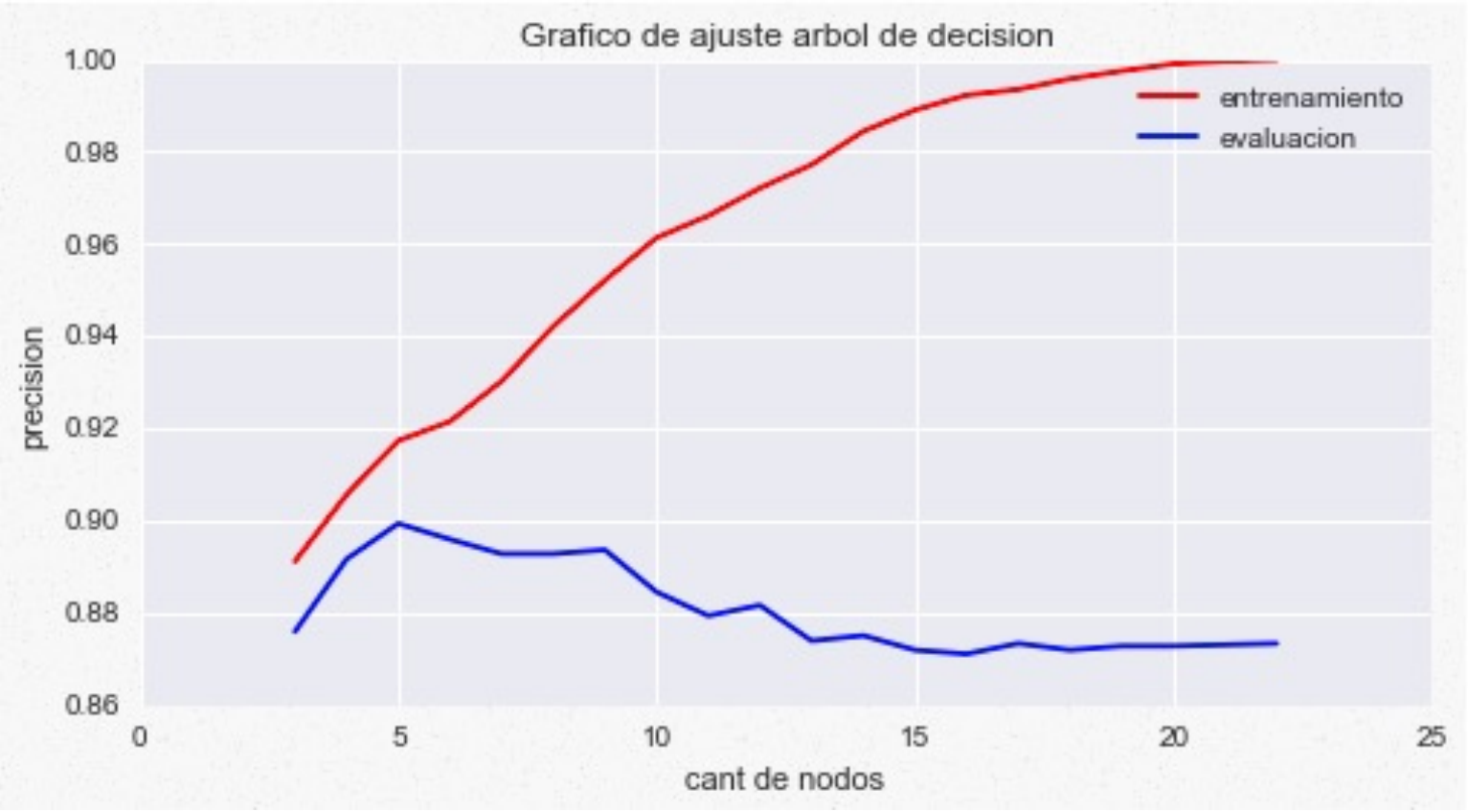
Modelo de regresión



Buen ajuste

Sobre

Sobreajuste <overfitting> = no generalización



Entre más complejos es el modelo, mayor sobre ajuste (línea roja) y menor capacidad de generalización (línea azul).

La complejidad del modelo en este caso está dado por los nodos del árbol

Evitar sobre ajuste <overfitting>

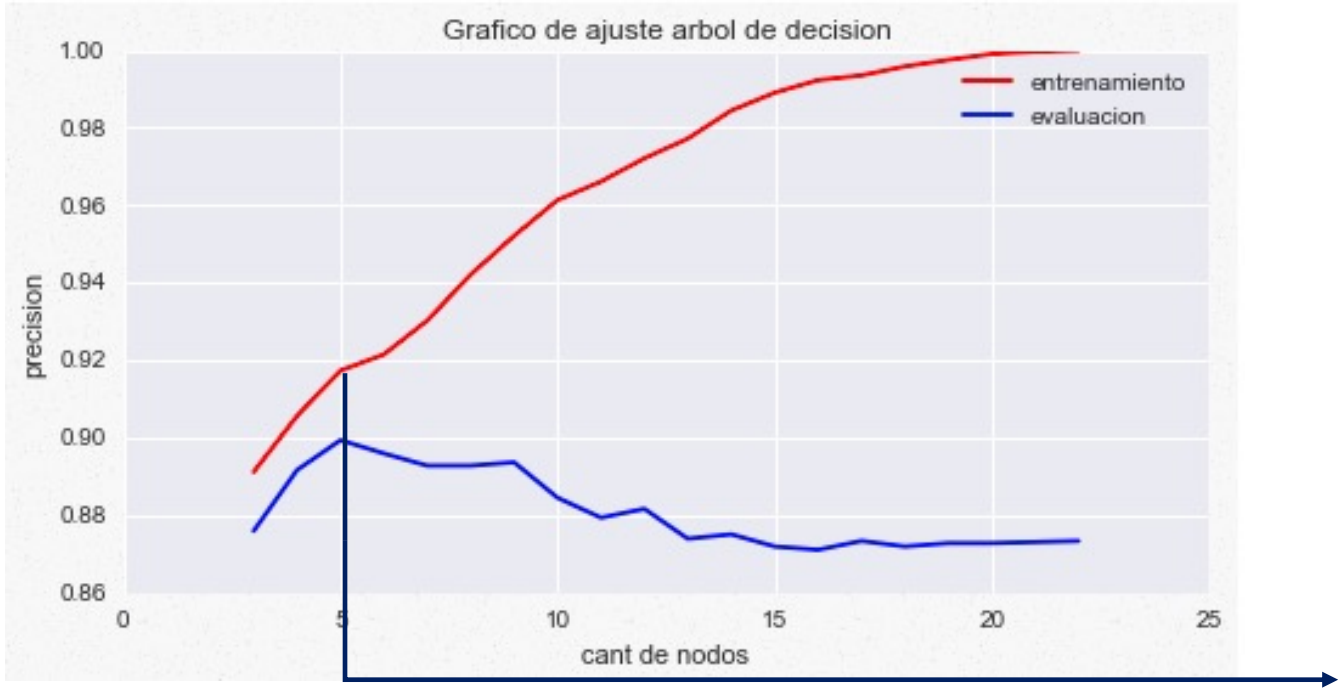
1. A partir de un conjunto de datos observados (donde se conoce su variable objetivo), dividir los datos en tres subconjuntos, denominados entrenamiento, validación y test.

Evitar sobre ajuste <overfitting>

2. Estimar distintos modelos (con distintos hiperparámetros) utilizando la data de entrenamiento, luego, aplicar estos modelos sobre los datos de validación.

Evitar sobre ajuste <overfitting>

3. Una vez aplicado los distintos modelos (por cada conjunto de hiperparámetros), se procede a calcular una métrica, luego, se debe seleccionar el conjunto de hiperparámetros de tal forma que el modelo generalice bien sobre los datos de validación.

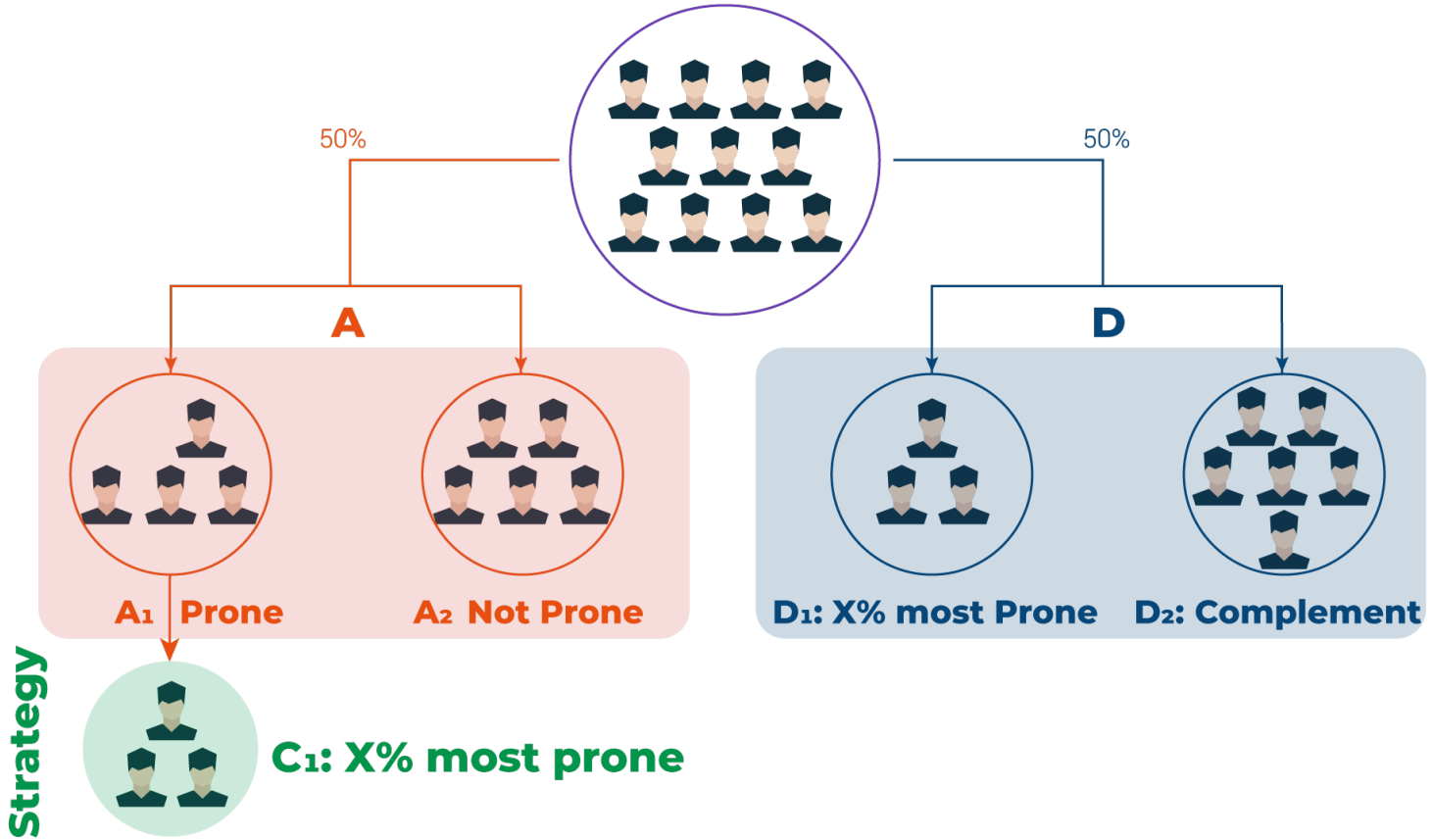


En este caso, la complejidad del modelo está dado por los nodos, cuando el árbol se estima con 5 nodos, se presenta una mejor capacidad de generalización

Evitar sobre ajuste <overfitting>

- GridSearch
- Validación cruzada
- Optimización bayesiana (Hyperopt, Optuna).

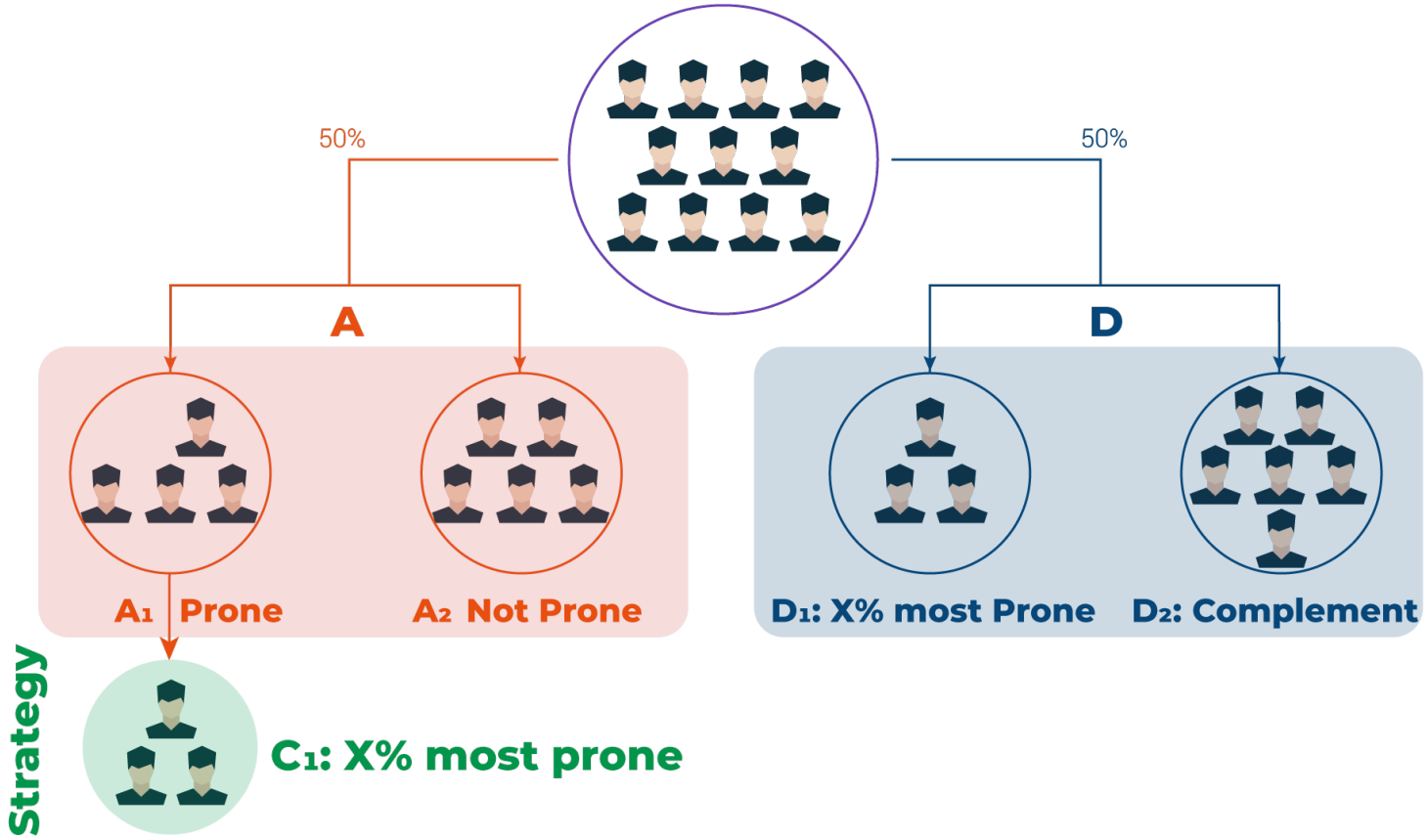
Se recomienda siempre medir el impacto de negocio -> experimento fuga de clientes



Resultados esperados:
A1 > A2
C1 < D1

A1 vs. A2 Model comparison
C1 vs. D1 Strategy comparison

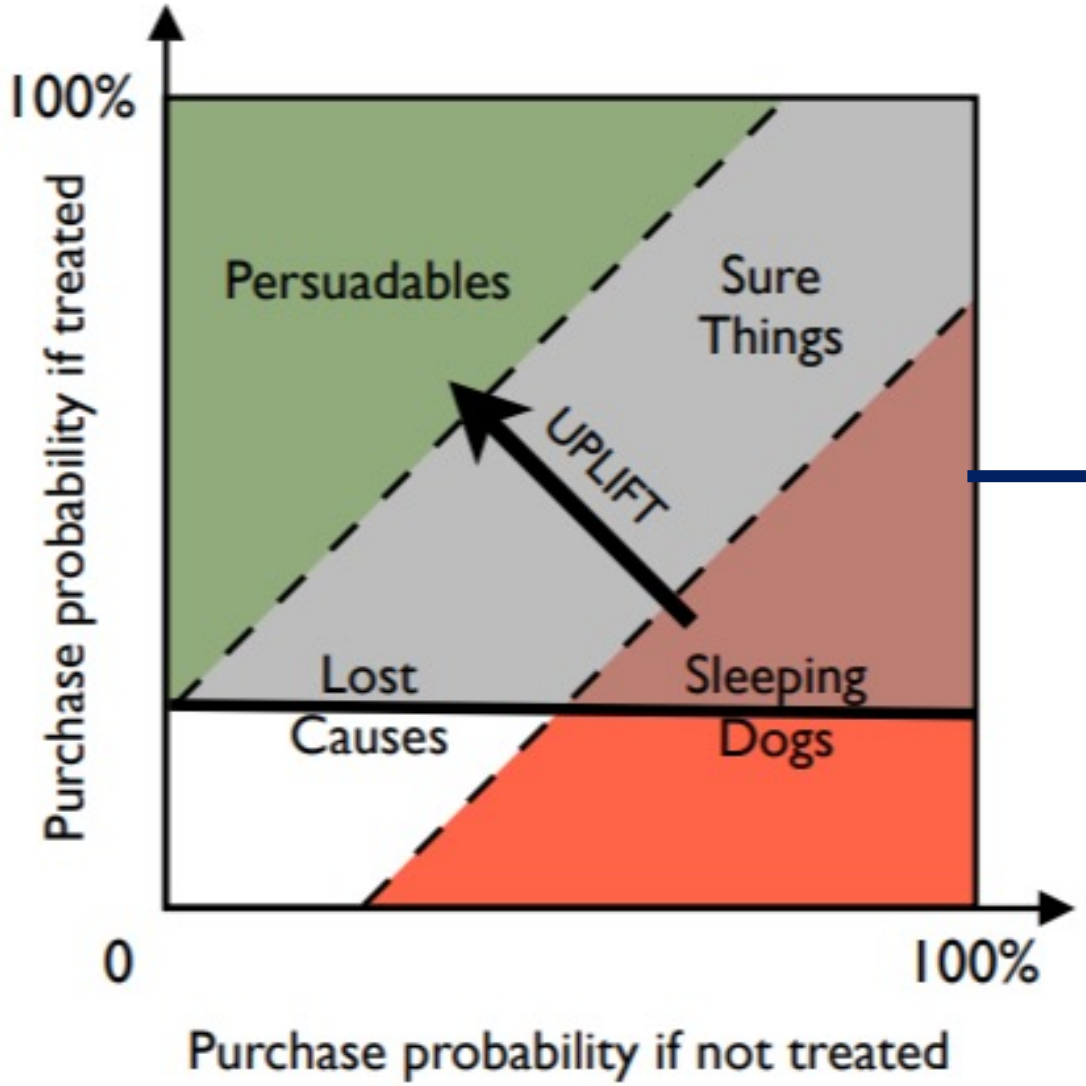
Algo muy común en la experimentación



Resultados frecuentes:
-El modelo identifica claramente los clientes que se van a fugar.
-La estrategia no es efectiva

A1 vs. A2 Model comparison
C1 vs. D1 Strategy comparison

¿El por qué sucede lo anterior?



Se gestionaron clientes de todo tipo (persuadables, sure things, sleeping dogs y algunos lost causes).

Modelos de Uplift

Objetivo: para cada individuo estimar una medida de ***uplift*** que se define como:

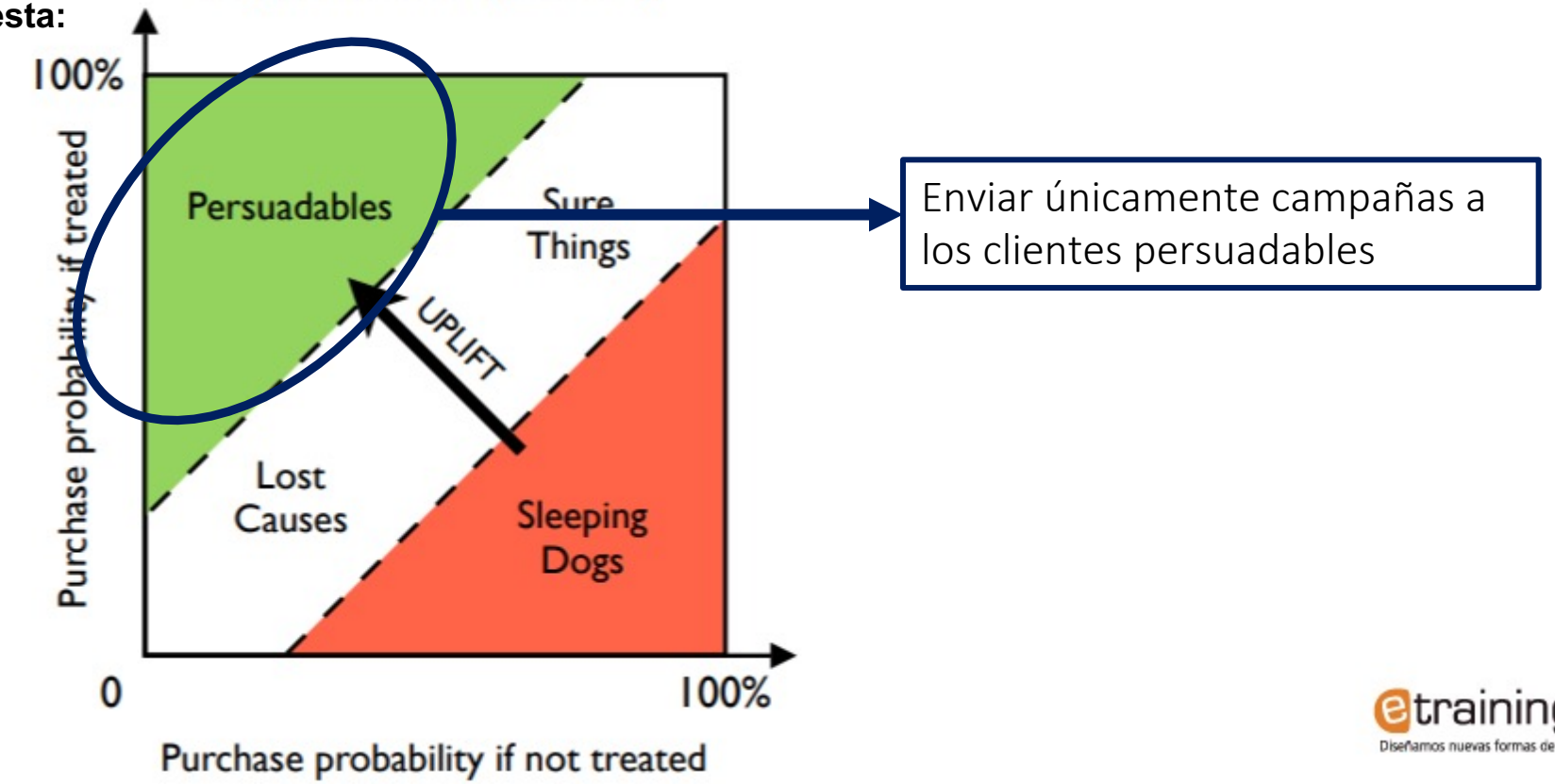
Uplift = probabilidad reacción dado oferta comercial – probabilidad reacción dado mercado natural

Retos:

Requiere datos de campañas anteriores.

Mayor complejidad estadística con relación a los modelos usuales de respuesta.

Propuesta:



Enviar únicamente campañas a los clientes persuadables

Cómo encontrar la estrategia adecuada para retener consumidores

Para generar estrategias de retención, es posible identificar grupos o segmentos de clientes con características similares, y por cada segmento se puede establecer una estrategia comercial distinta.

El futuro digital es de todos | MinTIC

Machine Learning <ML>

etraining Diseñamos nuevas formas de aprender | tecnalia Colombia Inspiring Business

Cómo encontrar la estrategia adecuada para retener clientes

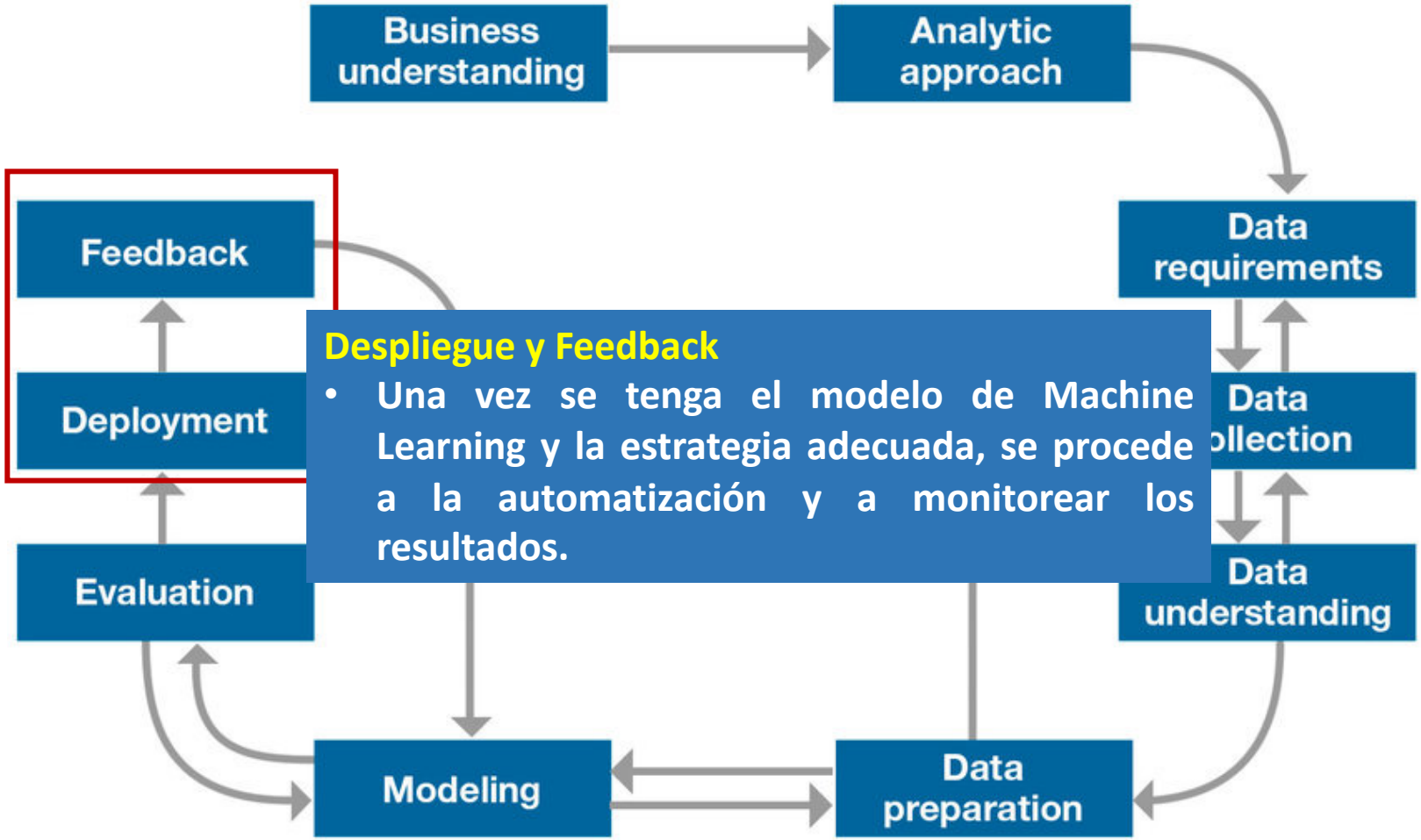
Para generar estrategias de retención, es posible identificar grupos o segmentos de clientes con características similares, y por cada segmento se puede establecer una estrategia comercial distinta.

Clientes que nunca han usado sus productos

Clientes con perfil digital y muy jóvenes

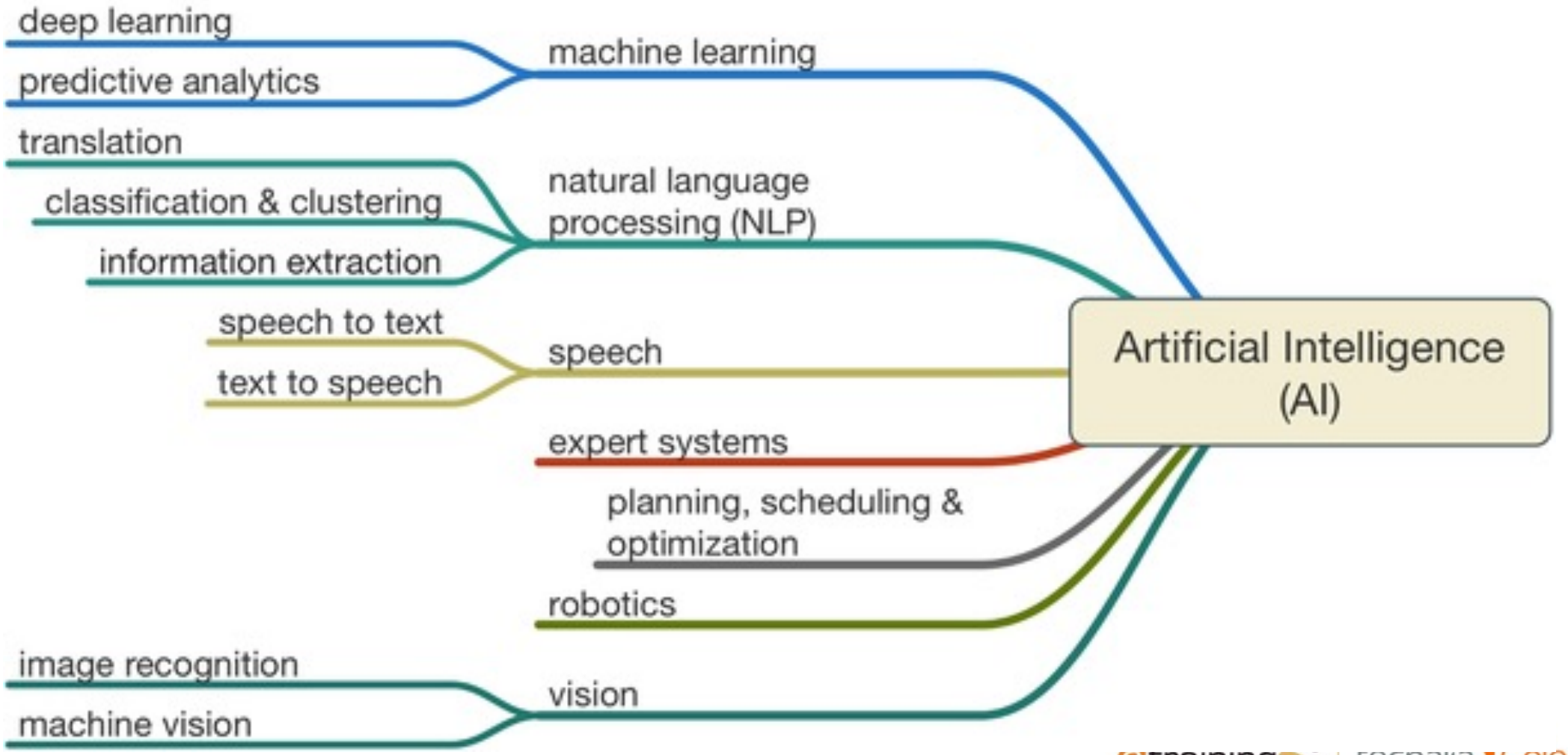
Clientes con perfil análogo y con experiencia

Machine Learning <ML>

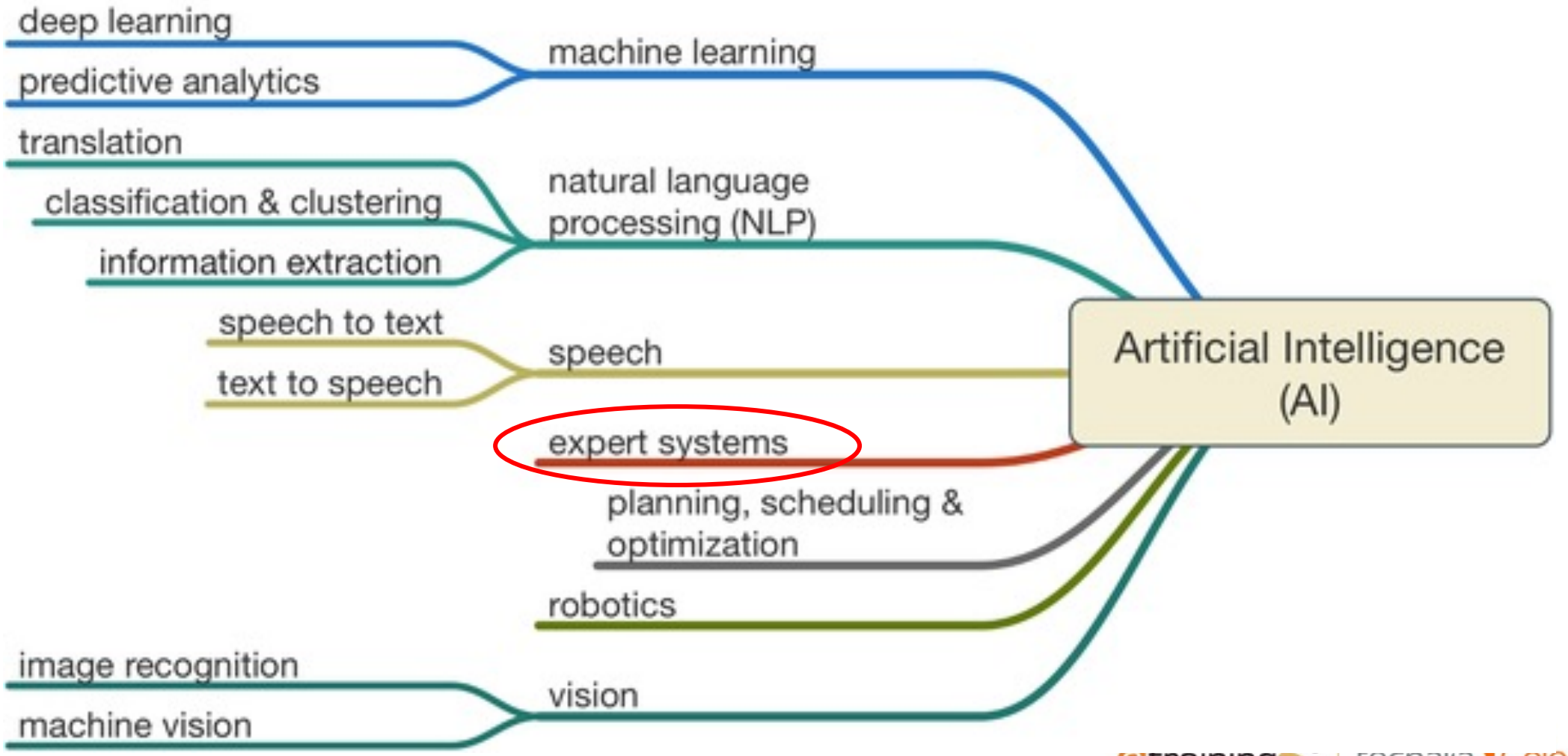


The IBM Foundational Methodology for Data Science. Source: [5].

Inteligencia artificial <AI>



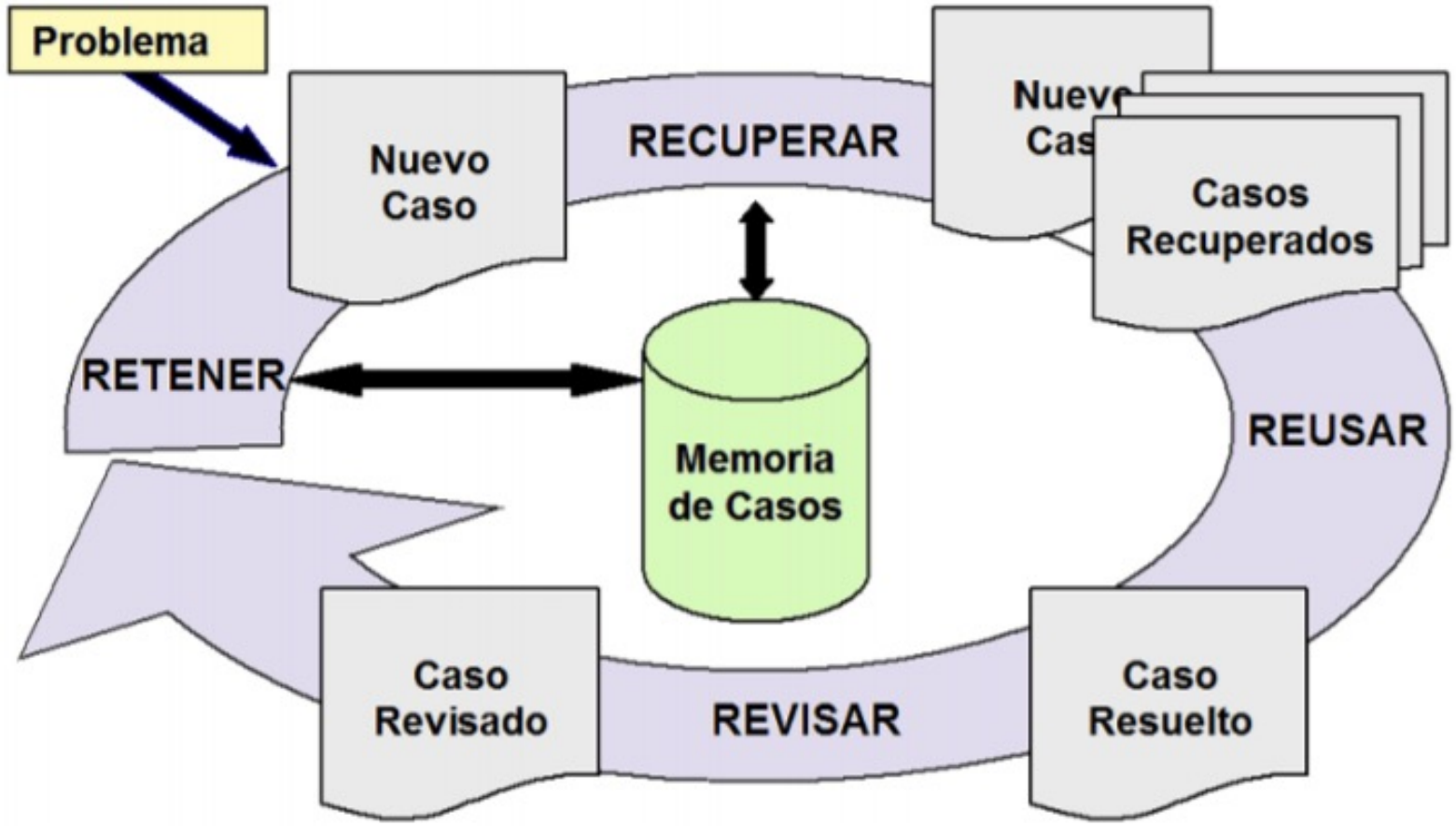
Inteligencia artificial <AI>



Sistema experto para resolver PQRS y atacar problemas de fuga de clientes

- Muchos clientes han cancelado sus productos por que cuando estos presentaron una queja o un reclamo, no se solucionó oportunamente. **-> Analítica prescriptiva**

Sistema experto para resolver PQRS y atacar problemas de fuga de clientes

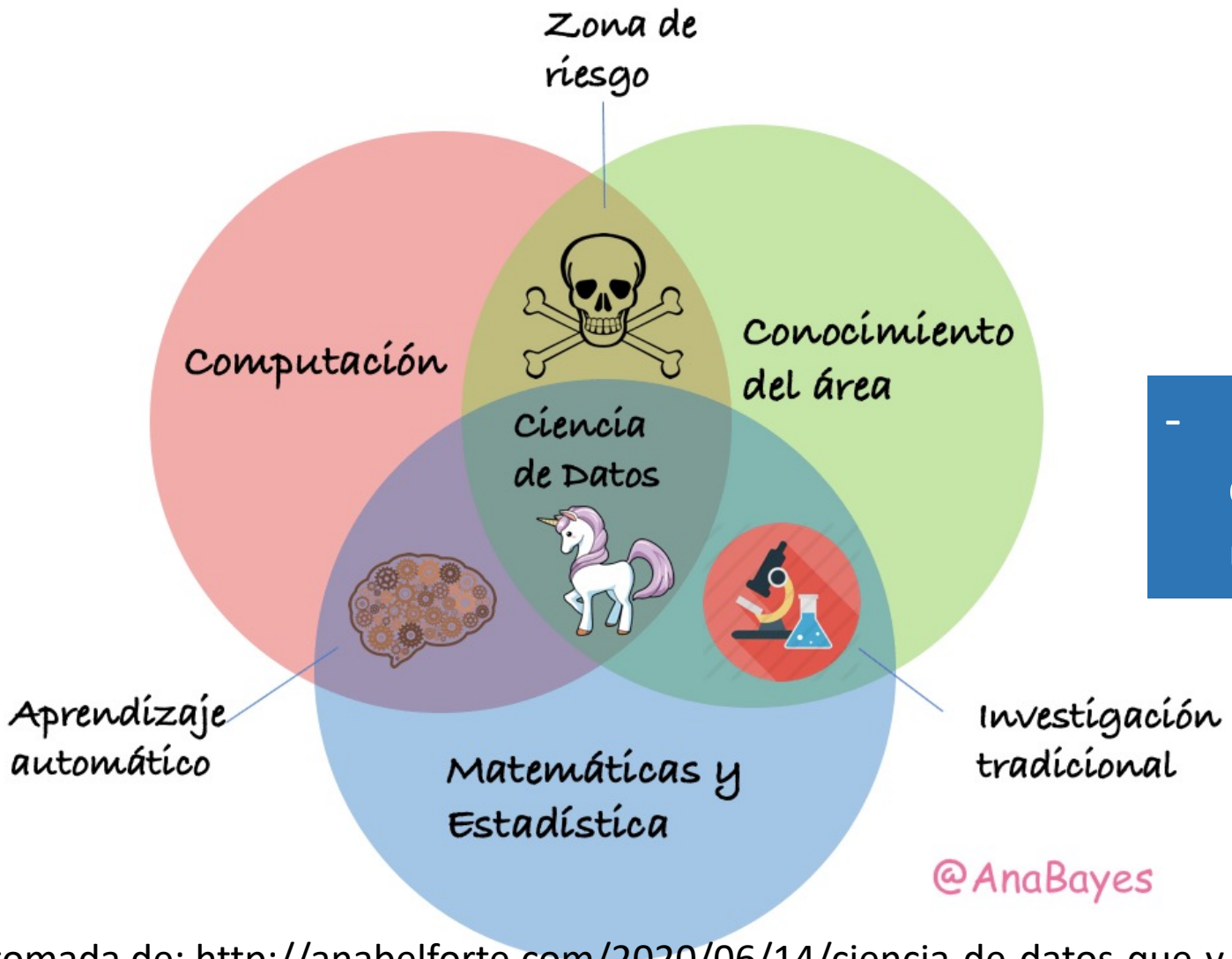


Problema	Solución
Problema 1	Solución 1
Problema 2	Solución 2

Dos impactos

1. Reducción tiempos solución PQRS, esto puede reducir la tasa de fuga de clientes
2. El conocimiento queda planteado en un sistema inteligente y no se “fuga” por rotación de personal.

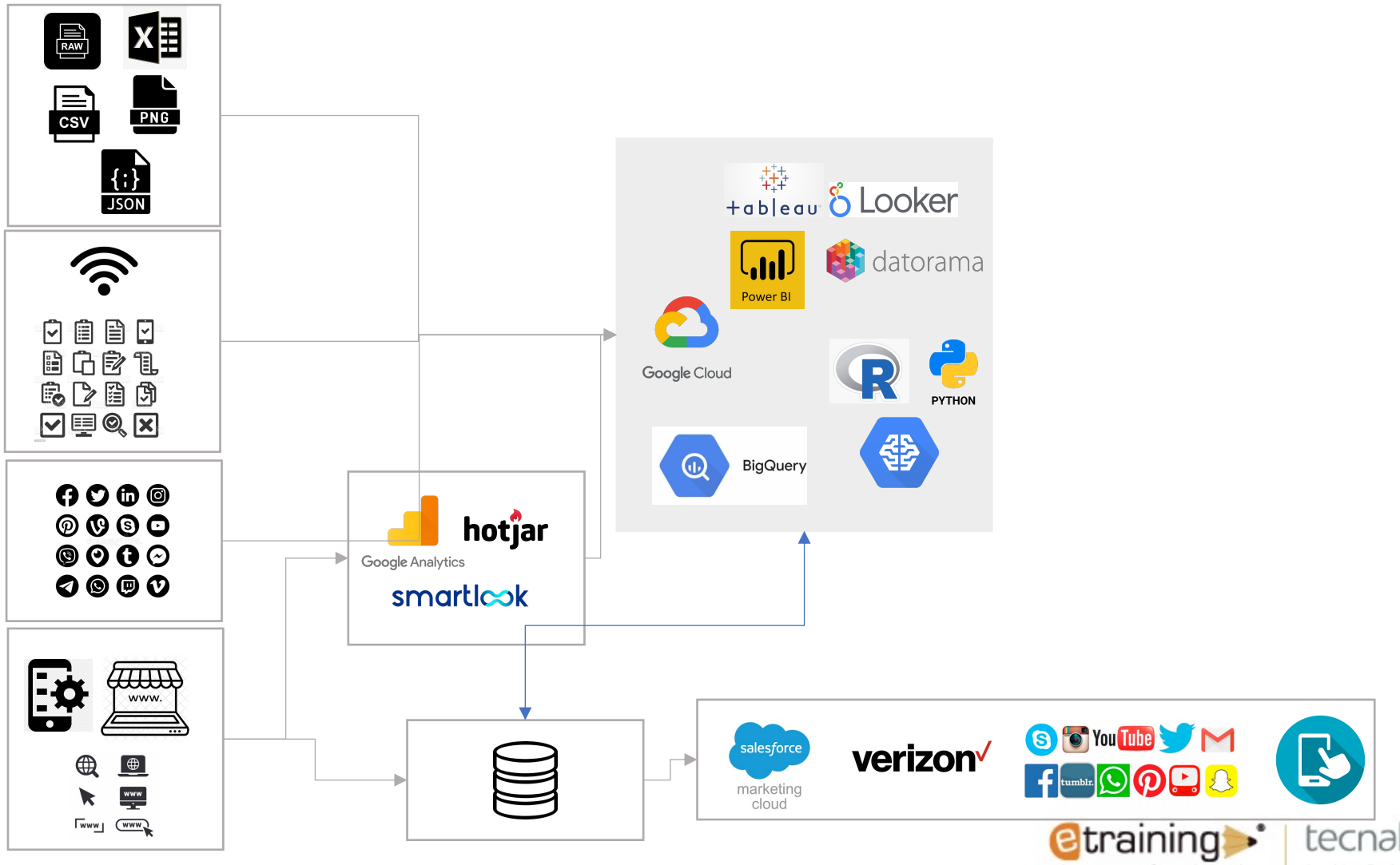
Retos: Perfiles unicornios



- Para construir soluciones basadas en datos, es importante contar con un equipo de trabajo.

Imagen tomada de: http://anabelforte.com/2020/06/14/ciencia-de-datos-que-y-por-que/

Retos: Los datos están pero no están



El futuro digital es de todos

MinTIC

Gracias

etraining
Diseñamos nuevas formas de aprender

tecnalia Colombia
Inspiring Business