Data Mining

Torture the data, and it will confess to anything.
— Ronald Coase, economist, Nobel Prize Laureate.

Definición

- Data mining es el proceso computacional de **explorar** y **descubrir patrones** en grandes datasets.
- Es un subcampo de las ciencias de la computación que combina variadas técnicas de la estadística, data science, teoría de bases de datos y aprendizaje automático.

Historia

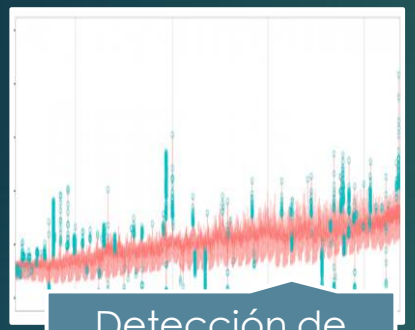
- Teorema de Bayes (1763)
- Regresión (1805)

- Turing (1936)
- Redes Neuronales (1943)
- Computación Evolutiva (1965)
- Bases de Datos (1970s)
- Algoritmos Genéticos (1975)

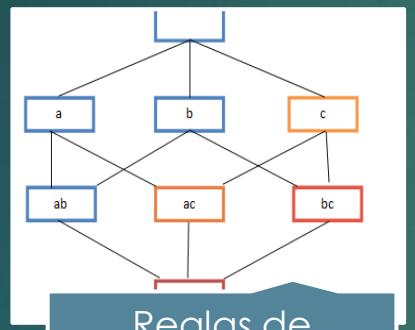
- KDD (1989)
- SVM (1992)
- Data Science (2001)
- Moneyball (2003)

- Big Data
- Adopción general
- DJ Patil (2015)

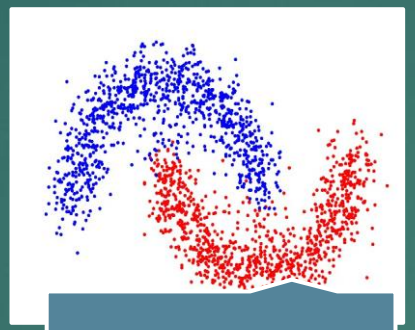
Usos Habituales



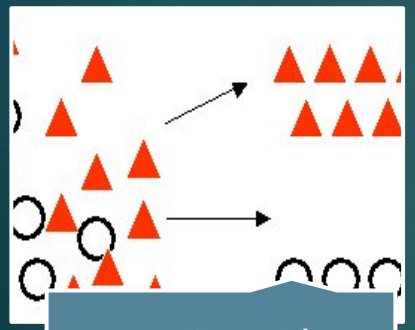
Detección de Anomalías



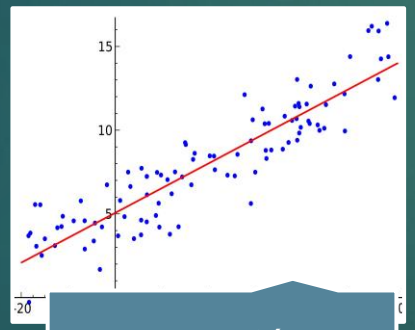
Reglas de Asociación



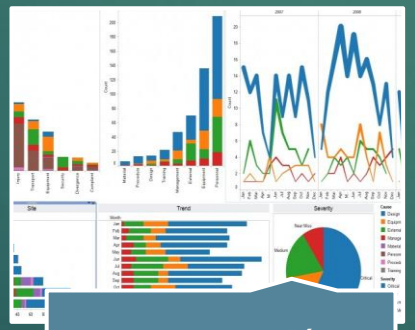
Clustering



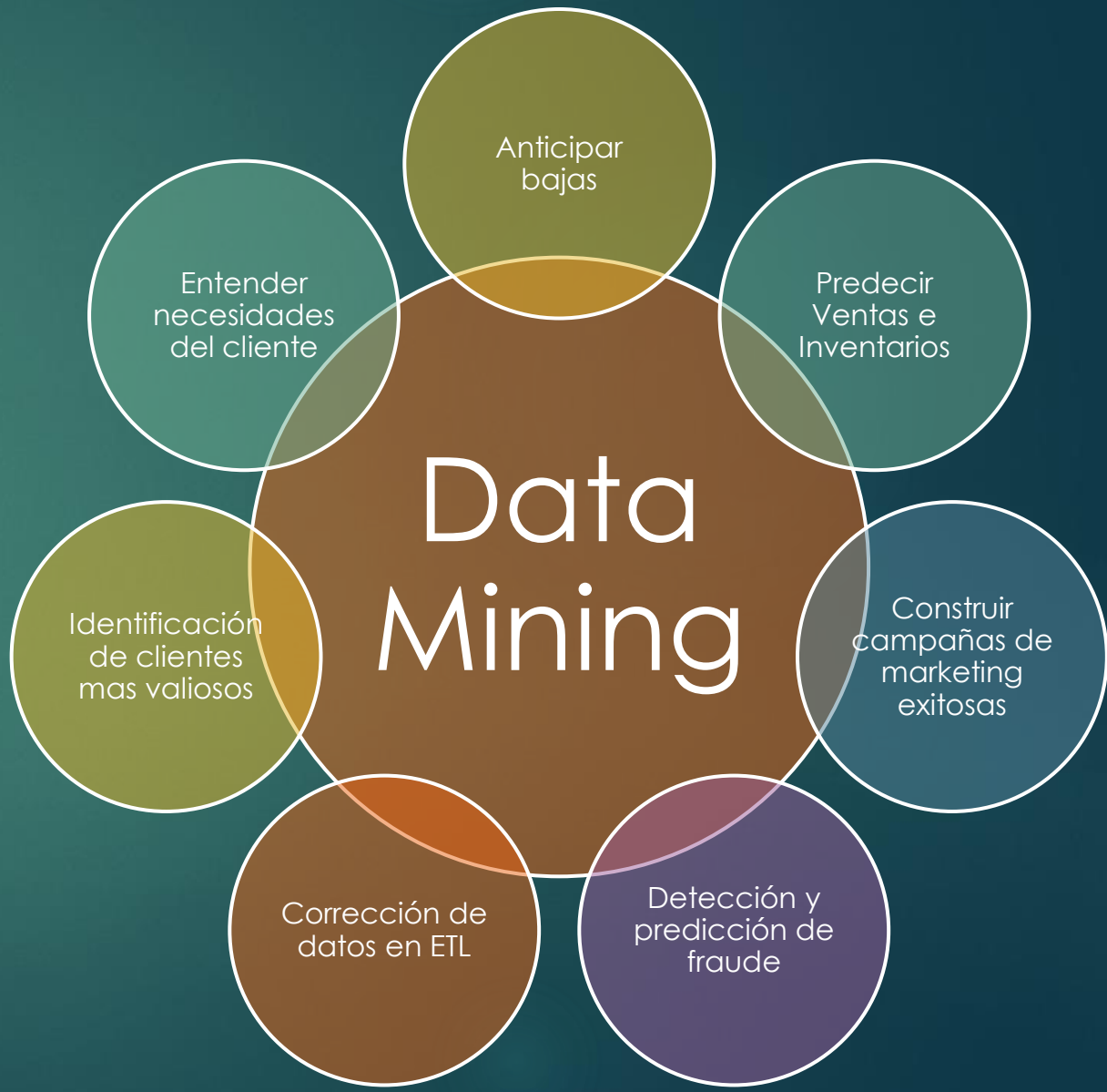
Clasificación



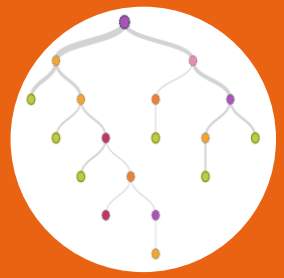
Regresión



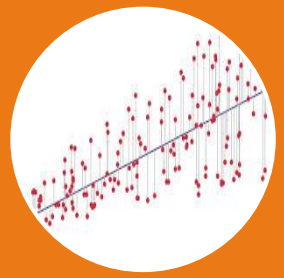
Sumarización



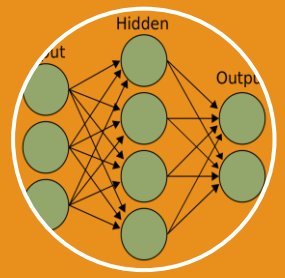
Algoritmos empleados



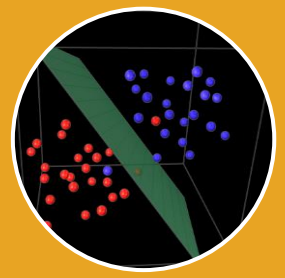
Arboles de Decisión



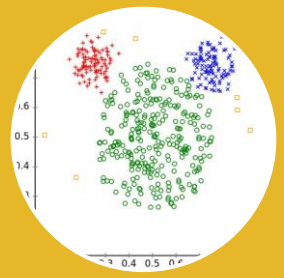
Regresión



Redes Neuronales



Support Vector Machine



K-means Clustering

Ensembles / Conjuntos

Para un mismo modelo se puede usar uno o más de un algoritmo. Esto puede realizarce en secuencia o como una combinación de modelos donde el ensemble (o combinación) puede superar el mejor de los modelos.


LANE CLOSED
LANE CLOSED
E-ZPass
E-ZPass
E-ZPass
E-ZPass
CASH

Proceso habitual

Modelo SEMMA

Nuevos datos son incorporados y el modelo se recalibra.

SAMPLE

Muestrear la información creando un data set lo suficientemente completo para incluir la información relevante.

EXPLORE

Explorar los datos buscando relaciones anticipadas, y tendencias y anomalías no anticipadas - para obtener un entendimiento profundo e ideas para el analisis posterior.

MODIFY

Modificar los valores observados por la creación, selección y transformación de variables. Puede mejorar la precisión del modelo.

MODEL

Modelar la información usando herramientas analíticas para encontrar la combinación de valores que certeramente predice un resultado final.

ASSESS

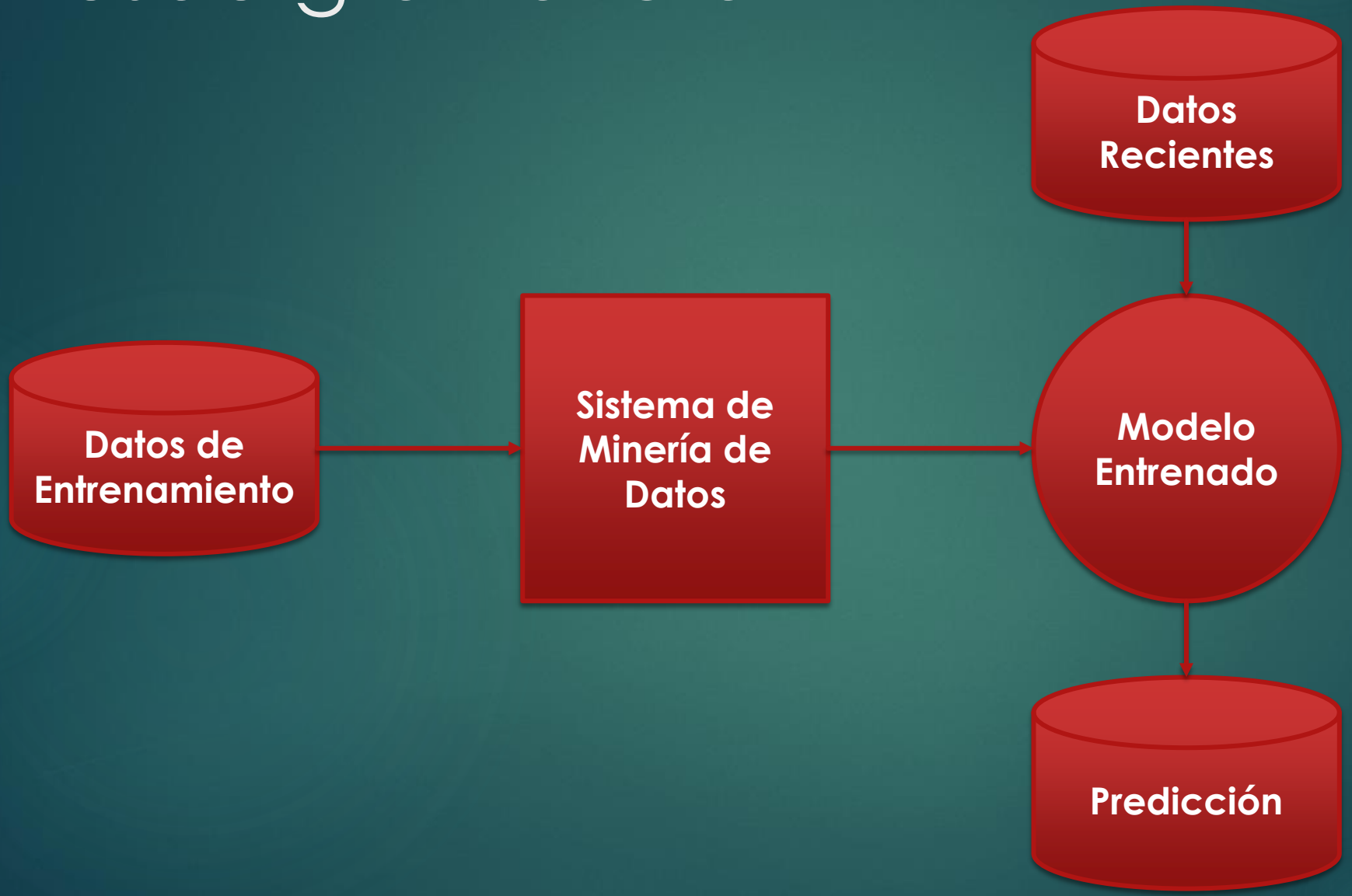
verificar y contrastar los modelos obtenidos para evaluar la usabilidad y confianza de los descubrimientos del proceso de data mining.

Retroalimentación

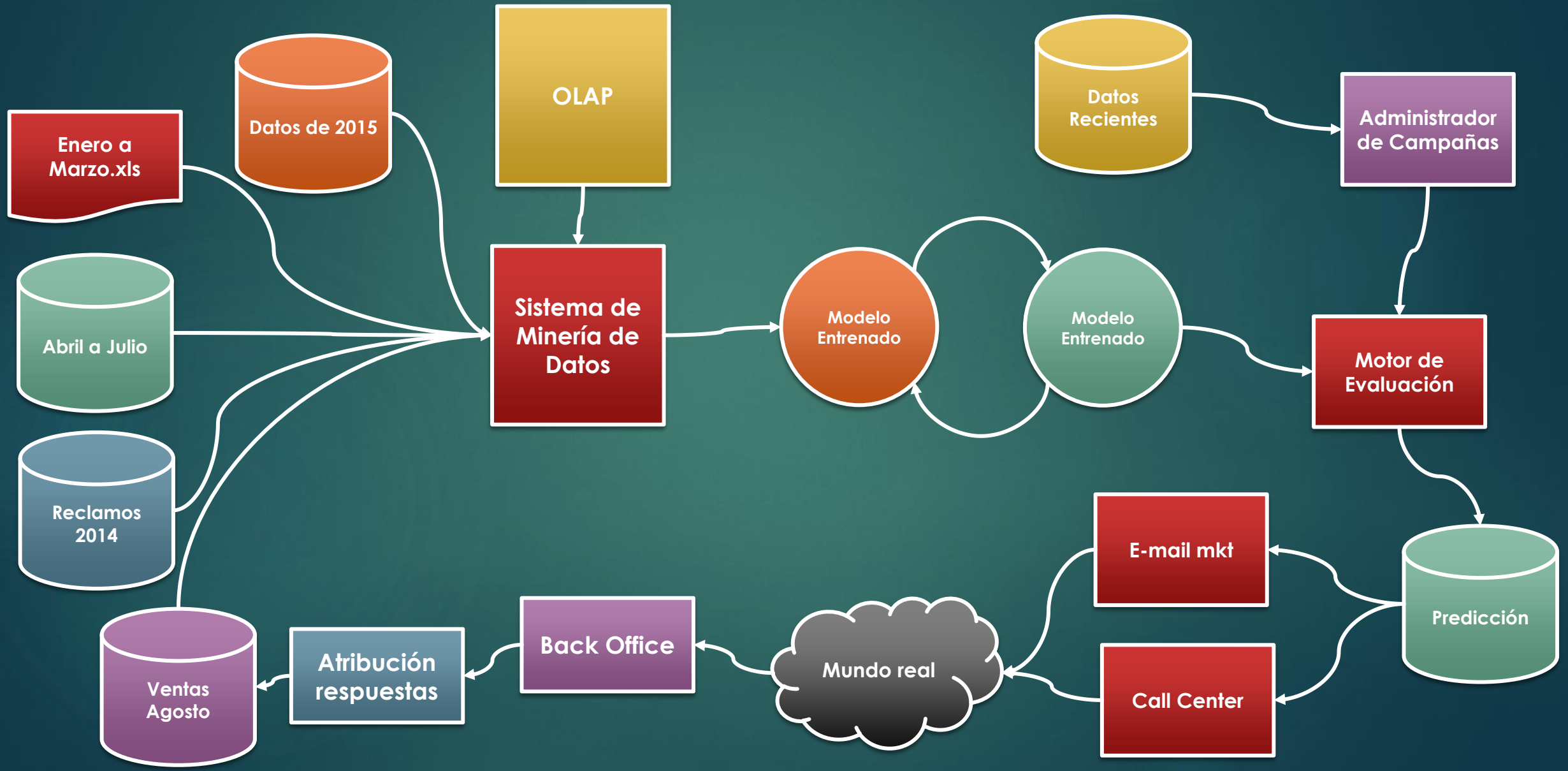
If we have data, let´s look at data. If all we have are opinions, let´s go with mine
— Jim Barksdale, former Netscape CEO.

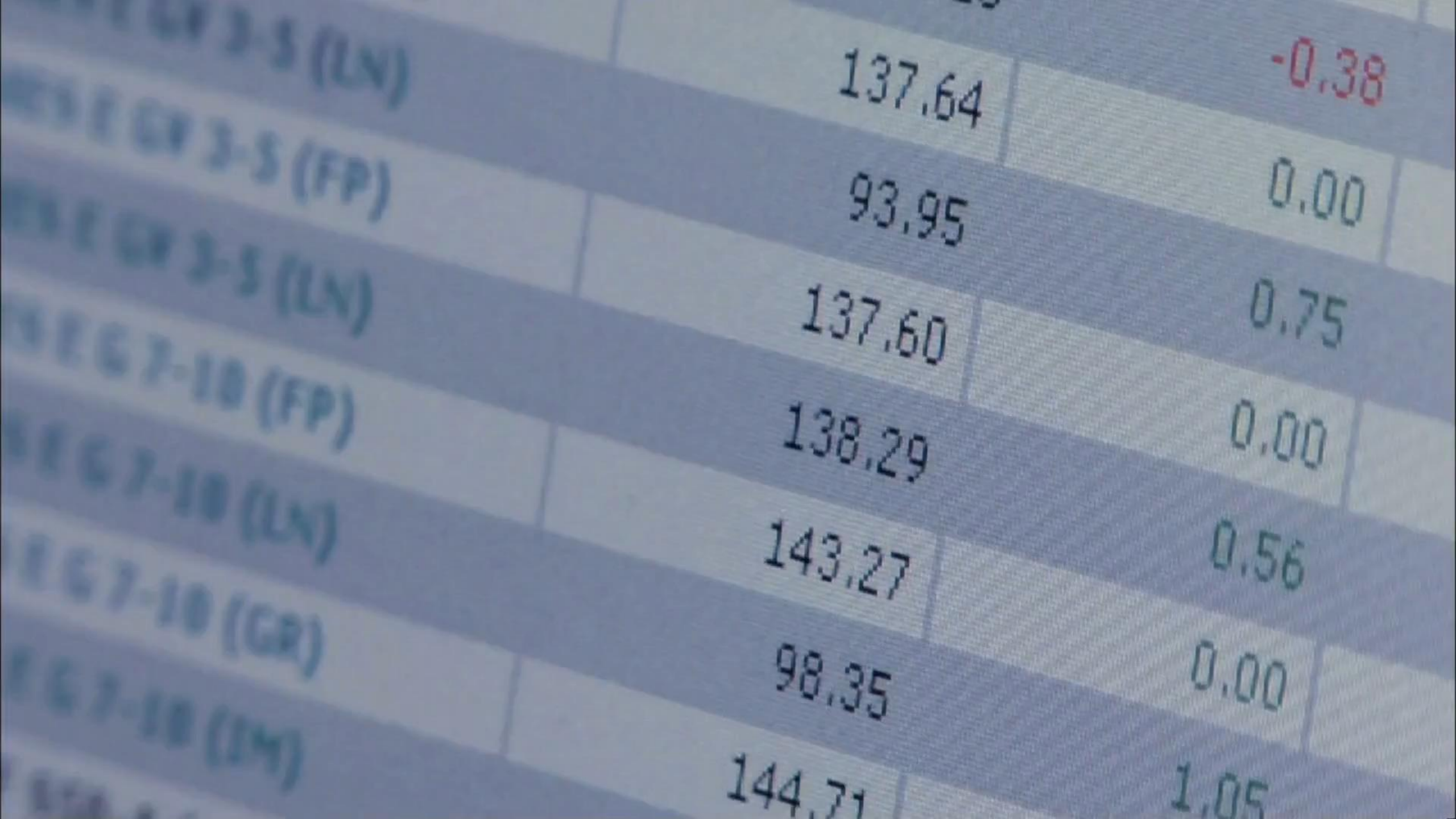
Cómo se construye y como se usa un modelo predictivo?

Desde gran altura



…y en el mundo real (cuando las cosas son simples)




137.64
-0.38
93.95
0.00
137.60
0.75
138.29
0.00
143.27
0.56
98.35
0.00
144.71
1.05

In God we trust, all others must bring data.
— W. Edwards Deming, engineer, statistician, professor, author, lecturer, and management consultant.

Como interpretar un modelo

Introducción – matriz de confusión

		Predicho	
		Positivo	Negativo
Real	Positivo	VP	FN
	Negativo	FP	VN

Un modelo de clasificación binaria produce 4 resultados:

- **Verdaderos Positivos:** Digo que se dan de baja y realmente se dan de baja.
- **Falsos Positivos:** Digo que se dan de baja y no lo hacen.
- **Falso Negativo:** Digo que continúan con la empresa y se dan de baja.
- **Verdadero Negativo:** Digo que continúan con la empresa y lo hacen.

Métricas: (Positivo = Baja; Negativo = Continúa)

- Exactitud:
  - Clasificaciones correctas / Casos $Ex = \frac{VP + VN}{VP + FN + FP + VN}$
- Precisión:
  - Indica proporción de bajas reales sobre bajas predichas $Pr = \frac{VP}{VP + FP}$
- Sensibilidad:
  - Indica proporción de las bajas reales que el modelo captura $S = \frac{VP}{VP + FN}$

Introducción – matriz de confusión

		Predicho	
		Positivo	Negativo
Real	Positivo	**VP**	**FN**
	Negativo	**FP**	**VN**

Exactitud	Precisión	Sensibilidad
$Ex = \frac{VP + VN}{VP + FN + FP + VN}$	$Pr = \frac{VP}{VP + FP}$	$S = \frac{VP}{VP + FN}$
VP FN / FP VN — VP FN / FP VN	VP FN / FP VN — VP FN / FP VN	VP FN / FP VN — VP FN / FP VN

Introducción – comparación de modelos

…pero…

…de acuerdo al problema me puede interesar una de las métricas más que las otras. Puedo ajustar el modelo para que se ajuste a este criterio.

El modelo me permite establecer el valor de corte (threshold).
Este será el grado de seguridad que le exijo antes de catalogar un caso como "positivo".

Salud:
Prefiero falsas alarmas a una lesión o muerte por falta de previsión.

Sensibilidad

Justicia:
…inocente hasta que se demuestre lo contrario.

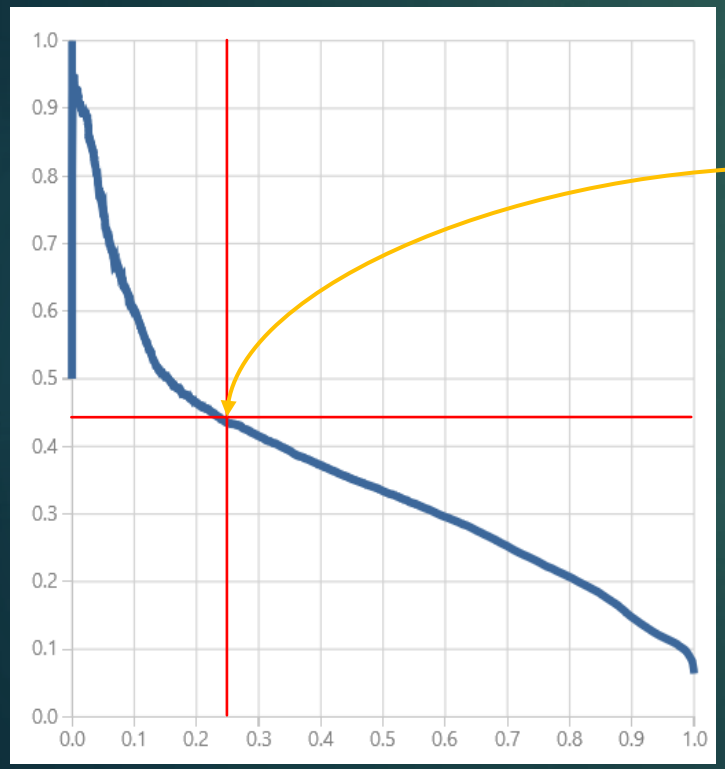
Precisión

Esto genera una matriz de confusión para cada valor de corte posible.

Resultados del modelo predictivo

Se presentan algunas maneras de visualizar la performance de un modelo predictivo.

Precisión - Sensibilidad



- Precisión: Proporción de bajas reales sobre bajas predichas.
- Sensibilidad: Proporción de bajas predichas sobre total de bajas.

Puedo capturar el 25% de las bajas acertando en el 45% de mis predicciones.

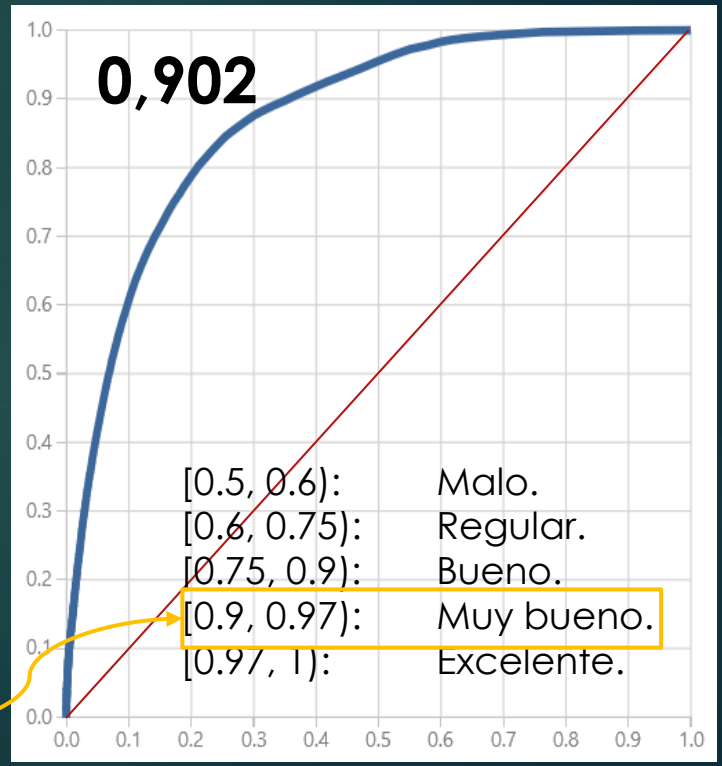
Resultado 8 veces superior al azar.

La curva ROC permite evaluar modelos predictivos.

El valor obtenido por el modelo es 0,902

Se puede considerar como un modelo Muy Bueno

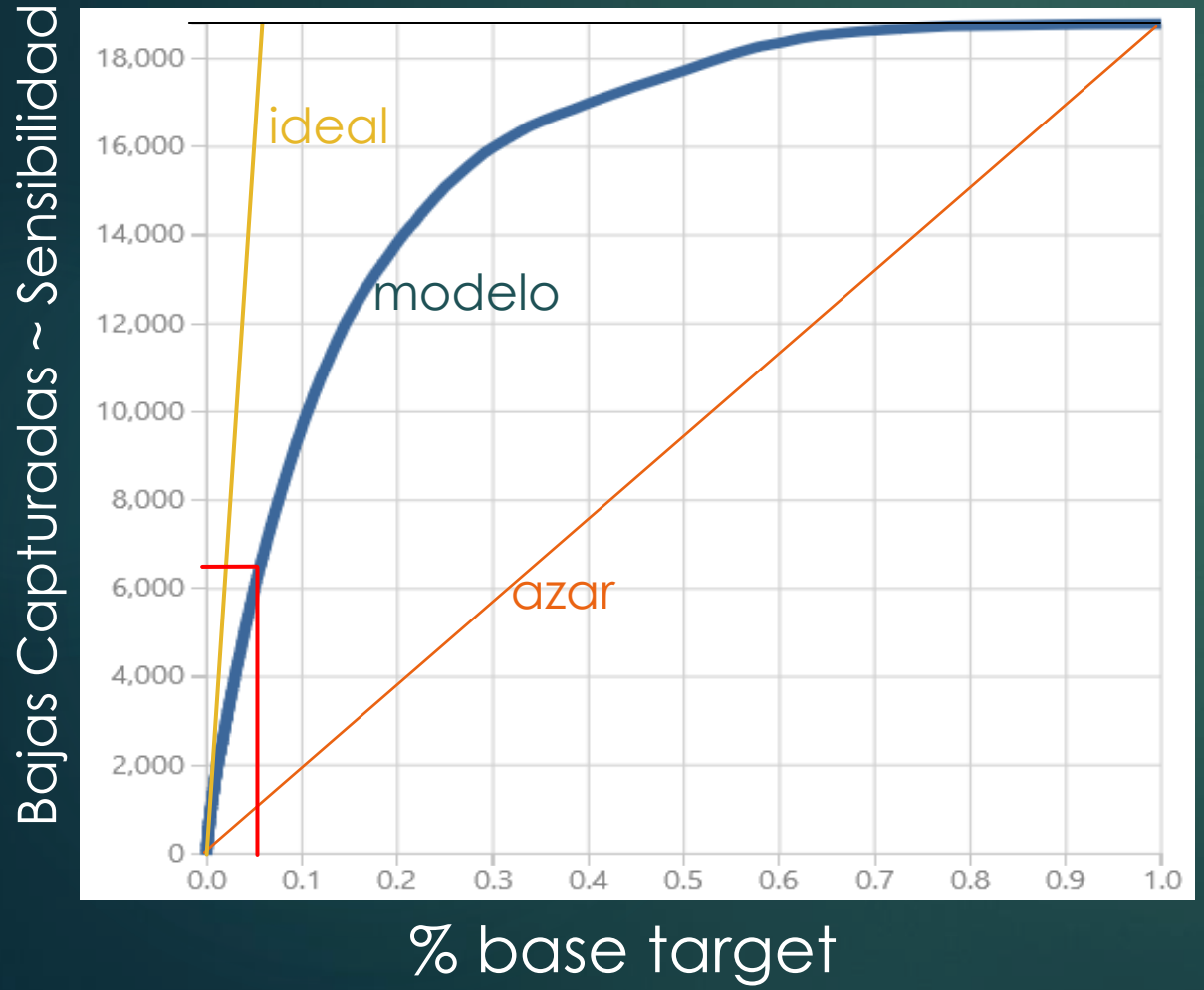
Curva ROC



Resultados del modelo predictivo

Otra manera de ver la performance es la Curva Lift

Curva Lift



La curva Lift muestra el número de bajas reales capturadas de acuerdo al número de clientes contactados en una acción de retención.

En este caso observamos que con este modelo contactando al **5%** de la base (15.000 de 300.000 clientes) se obtiene un **32%** de las bajas reales (6.100 de 19.200 bajas del período).

Para conseguir esa cantidad de bajas en una muestra aleatoria hubiera sido necesario contactar 95.300 clientes.

El modelo define una curva, pero somos nosotros los que debemos definir un punto específico en la curva.

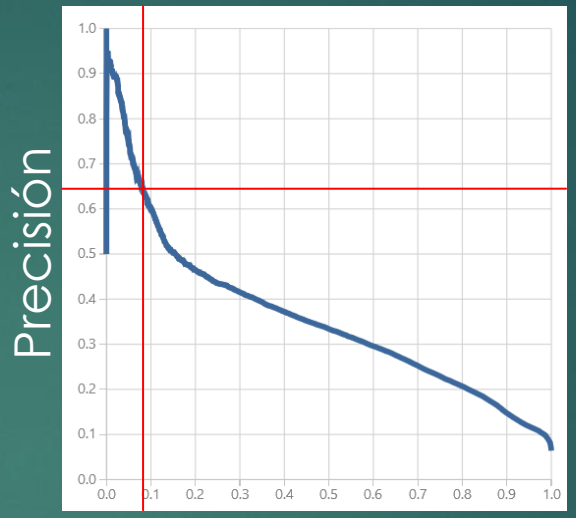
Resultados con distintos valores de corte

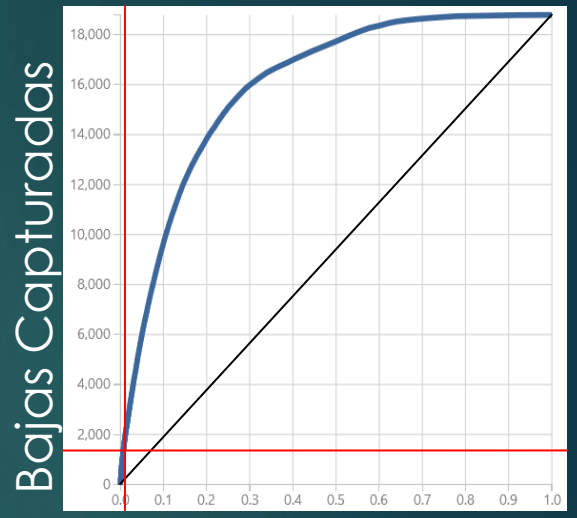
Modelo exigente

Precisión vs azar = 10

Ratio bajas = 6,4%

		Predicho	
		Positivo	Negativo
Real	Positivo	**1586**	**17593**
	Negativo	**906**	**279915**





Apunto al 0,8% de mi base y obtengo el 8,3% de las bajas

Exactitud	Precisión	Sensibilidad
94%	64% **10 veces superior al azar**	8,3%

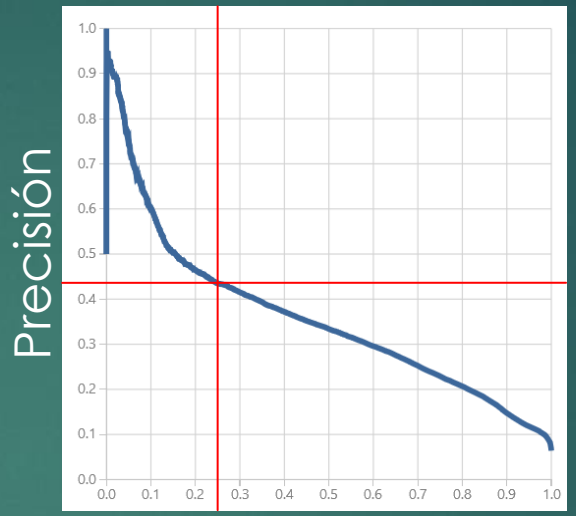
Resultados con distintos valores de corte

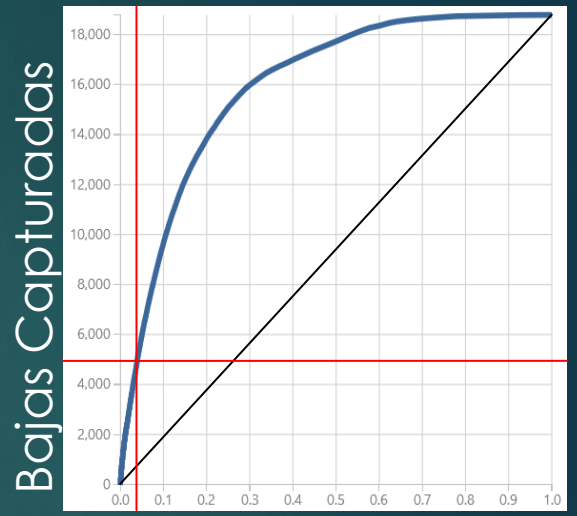
Modelo medio

Precisión vs azar = 7

Ratio bajas = 6,4%

		Predicho	
		Positivo	Negativo
Real	Positivo	**4845**	**14334**
	Negativo	**6310**	**274511**





Apunto al 3,7% de mi base y obtengo el 25,3% de las bajas

Exactitud	Precisión	Sensibilidad
93%	43% **7 veces superior al azar**	25%

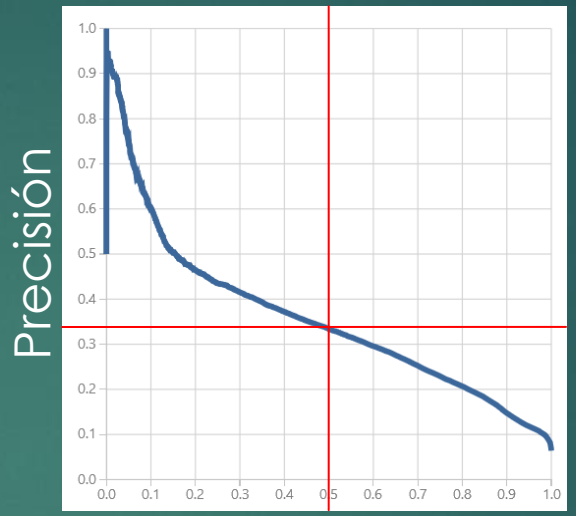
Resultados con distintos valores de corte

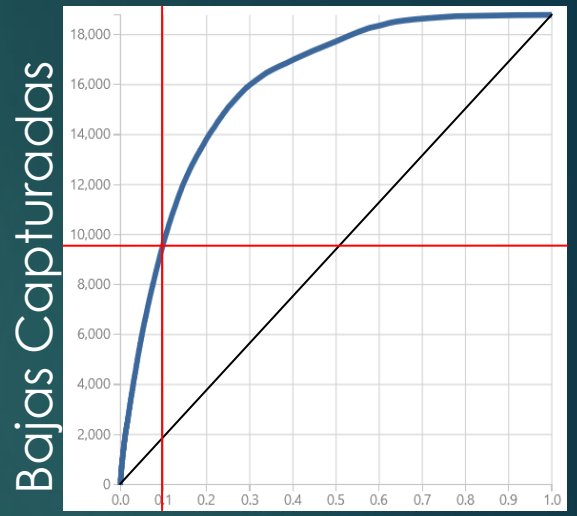
Modelo flexible

Precisión vs azar = 5

Ratio bajas = 6,4%

		Predicho	
		Positivo	Negativo
Real	Positivo	**9525**	**9654**
	Negativo	**18871**	**261950**





Apunto al 9,5% de mi base y obtengo el 50% de las bajas

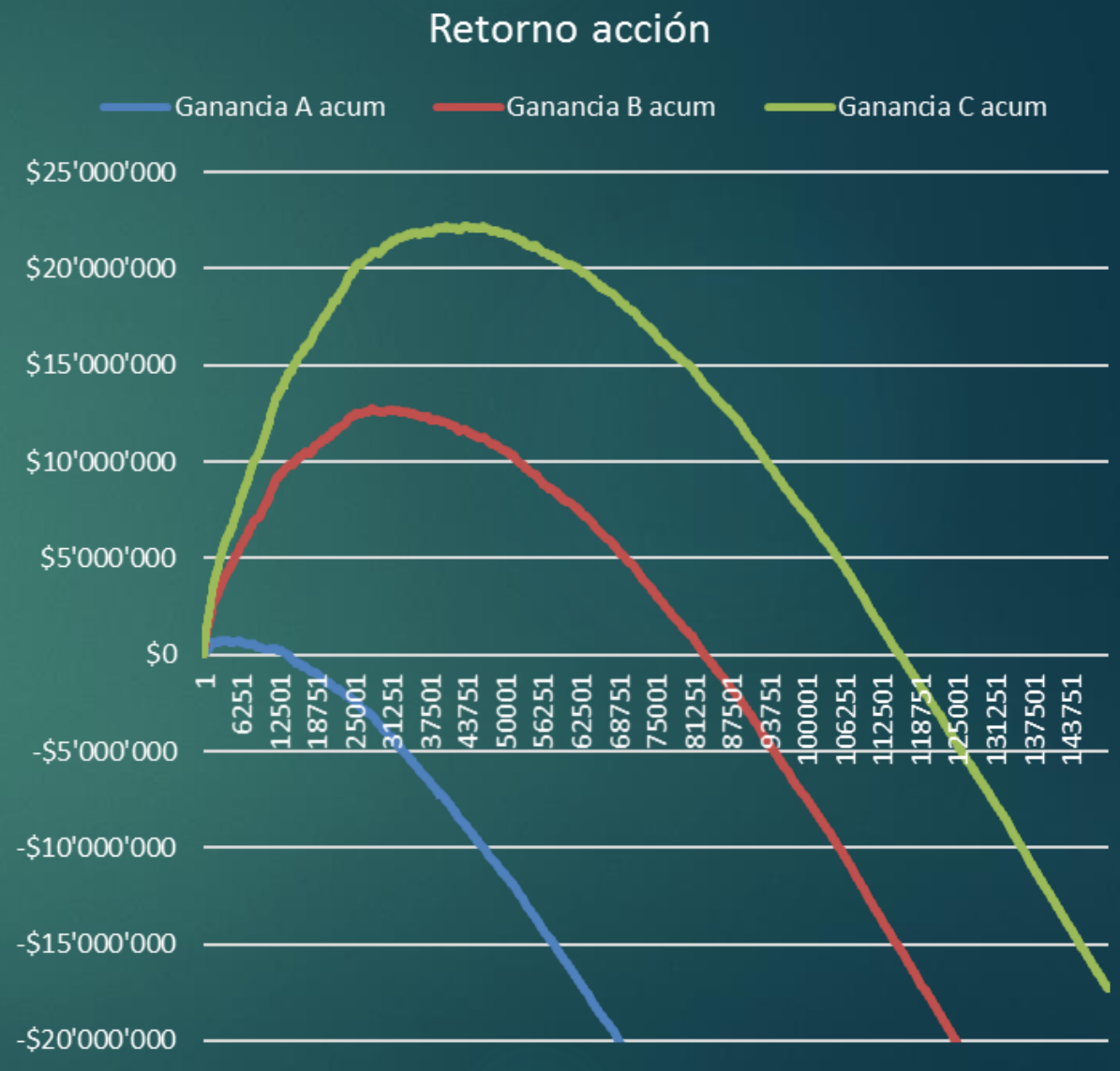
Exactitud	Precisión	Sensibilidad
90%	34% **5 veces superior al azar**	49,7%

Ejercicio Económico Supuestos

Inversión por Cliente	% Retención	% Gasto Bajas no retenidas	Ingreso Acierto	Clientes contactados	Bajas reales	Bajas evitadas	Inversión	Ingresos	Ganancia
• Costo del descuento o regalo a cada cliente incluído en la acción	• Porcentaje de las bajas que permanecen en la empresa por el incentivo de la acción de retención • Se podrá obtener con muestras pequeñas sujetas a diferentes incentivos.	• Si el incentivo se distribuye a lo largo de varios meses, los clientes que no se logró retener no recibirán el total del incentivo, y por lo tanto el costo será menor	• Ingresos adicionales para la empresa por cada cliente retenido	• Cantidad de clientes a los que se ofrece el incentivo	• Bajas que hubieran abandonado la empresa dentro del grupo de Clientes contactados.	• Bajas que se evitaron por la acción	• Monto que se distribuyó a todos los clientes a los que se les ofreció el incentivo	• Monto que ingresa a la empresa gracias a las Bajas evitadas	• Diferencia entre los ingresos y el costo de los incentivos (Ingresos – Inversión)

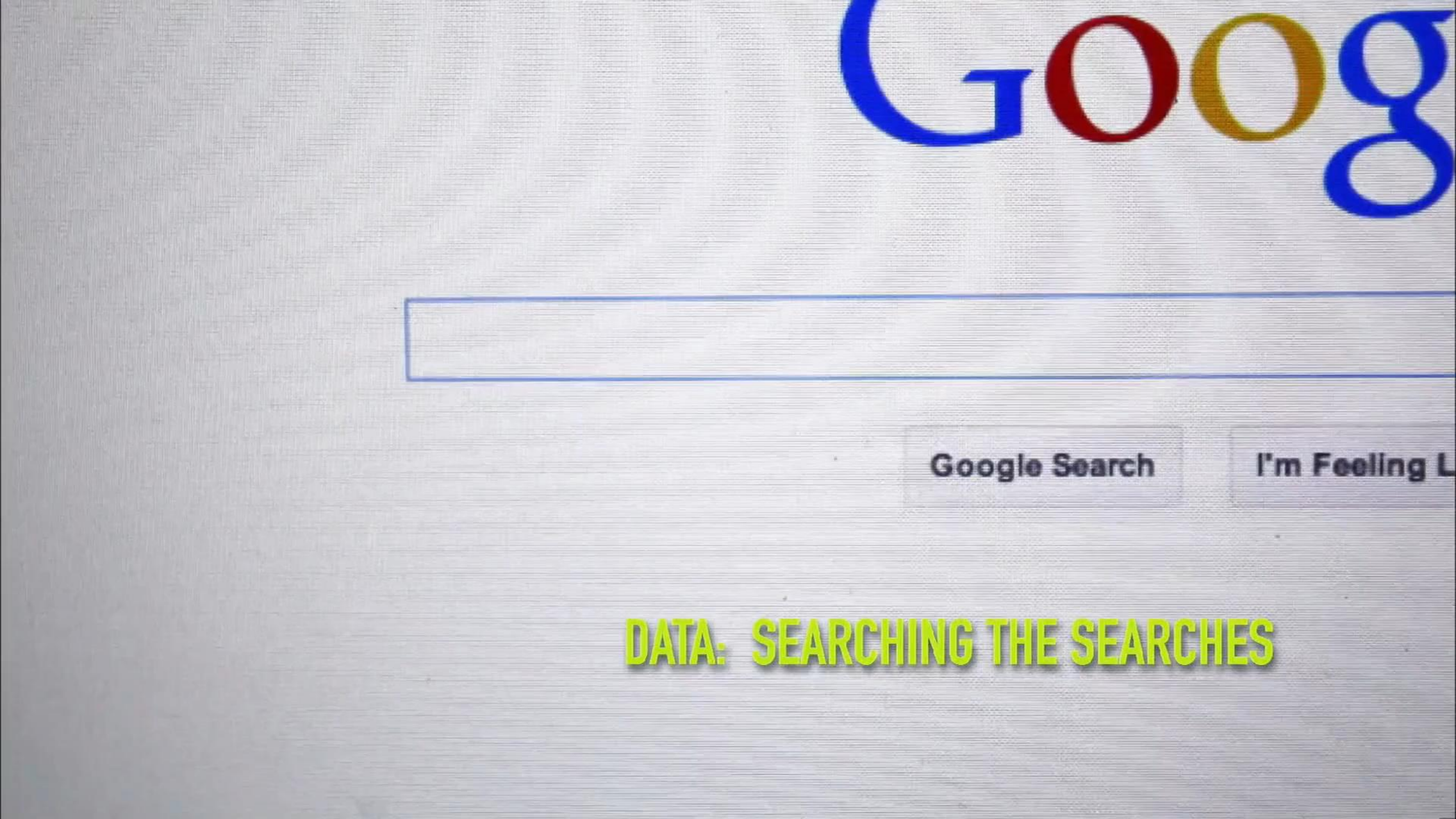
Ejercicio Económico

	Escenario A	Escenario B	Escenario C
Inversión por cliente	$600	$600	$600
% Retención	10%	20%	25%
% Gasto Bajas no retenidas	100%	66%	33%
Ingreso Acierto	$ 15'000	$ 15'000	$ 15'000
Clientes contactados	3791	27933	43241
Bajas reales	2164	9434	12212
Bajas evitadas	216	1887	3053
Inversión total	$ 2'475'600	$ 15'542'724	$ 23'532'708
Ingreso	$ 3'246'000	$ 28'302'000	$ 45'795'000
Ganancia	$ 770'400	$ 12'759'276	$ 22'262'292



Suponiendo que el descuento implica un costo de $600, que solo el 20% de las personas a las que se les ofrece el descuento deciden permanecer y que cada uno de estos genera luego $15.000; se evitarían 1887 bajas con un resultado neto de $12.000.000

Es posible (y recomendable) iniciar con muestras más pequeñas para validar los supuestos.

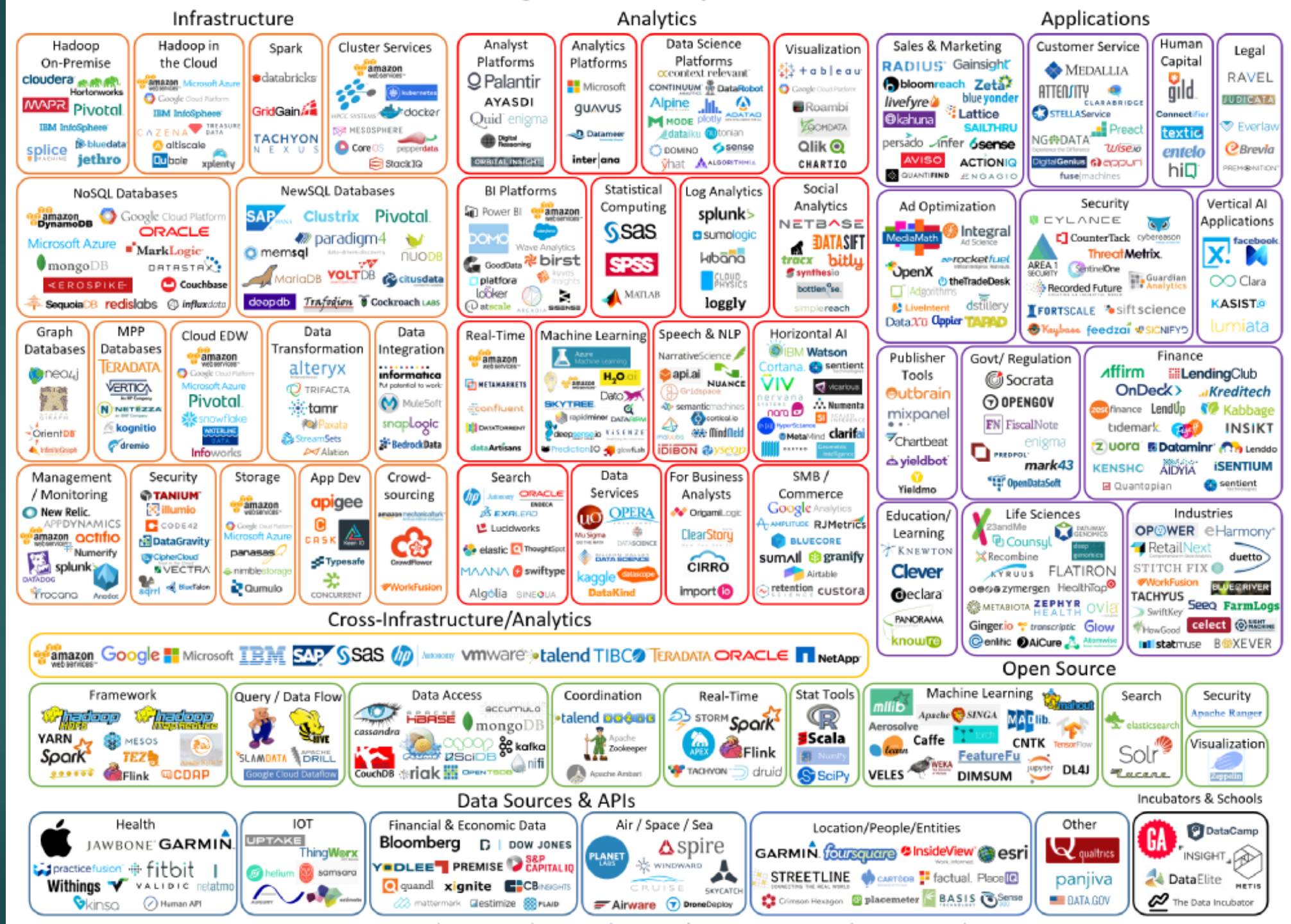

Goog
Google Search
I'm Feeling L
DATA: SEARCHING THE SEARCHES

Herramientas Disponibles

“Data!data!data!" he cried impatiently. "I can't make bricks without clay.”
— Arthur Conan Doyle, The Adventure of the Copper Beeches (Sherlock Holmes)

Landscape Completo

Big Data



Algunos productos destacados

Open Source

R
- www.r-project.org

Python
- www.python.org
- www.continuum.io

Weka
- www.cs.waikato.ac.nz/ml/weka/

KNIME
- www.knime.org

RapidMiner
- rapidminer.com

Pagas

SAS
- www.sas.com

SPSS
- www.ibm.com/analytics/

Cloud

Google Cloud
- cloud.google.com/ml/

Azure ML
- studio.azureml.net

Amazon
- aws.amazon.com/es/machine-learning/

Muchas gracias