

■ Primera entrega

- Una semana.
- Usar sólo Excel para las cuentas.
- Tareas:
 - Entendimiento de los datos y negocio.
 - Observar las distribuciones de las variables. Valores faltantes y outliers.
 - Buscar y eliminar variables correlacionadas.
 - Analizar las proporciones de churn para distintas variables.

■ Segunda entrega

- Una semana.
- Usar cualquier software para modelar y validar.
- Tareas:
 - Particionar los datos.
 - Generar tres modelos de árbol.
 - Evaluar los modelos.
 - Aplicar el modelo de predicción elegido (justificar cual y por qué) a la base de clientes y generar un score de churn para cada cliente.

■ Documentos:

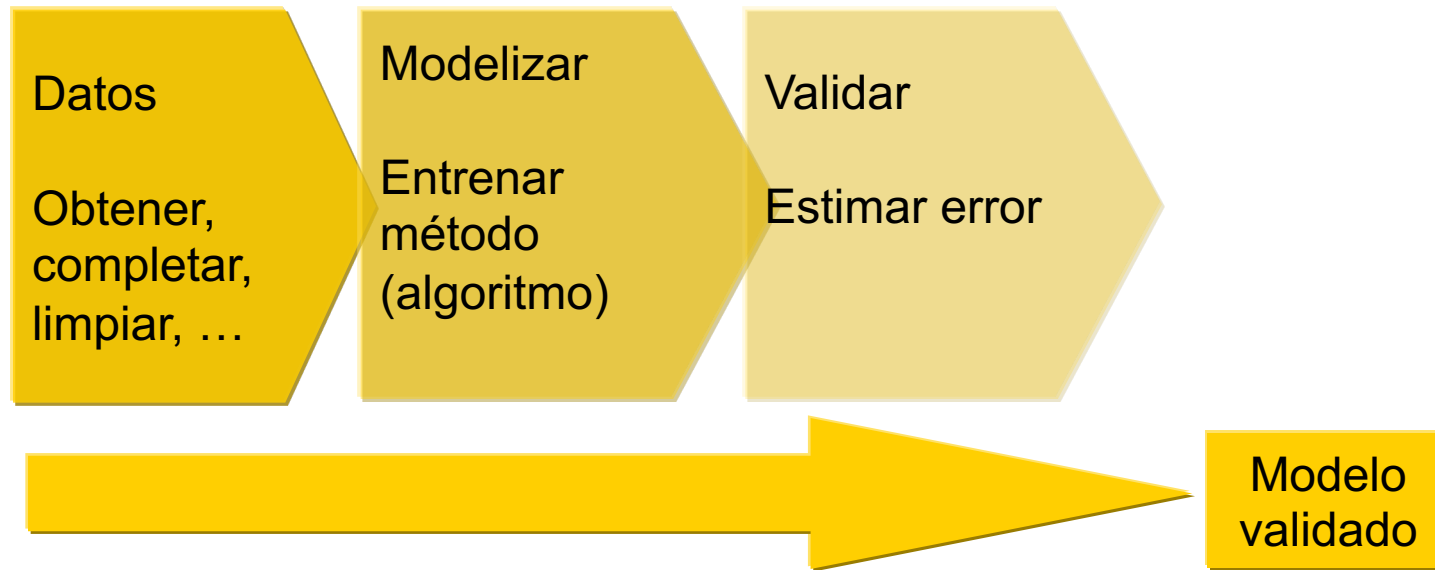
- Caso_2_churn.pdf: Cuestionario sobre entendimiento del negocio.
- TP_Telco.pdf: Descripción del dataset y ejemplos de descripción de variables.
- DATA MINING-TELCO.pdf: Conceptos sobre Churn.
- dm churn telecomunicaciones.pdf: texto sobre DM en fuga de clientes en Telcos.

■ Otros recursos:

- Esta presentación.
- XLSX, arff.
- 2 TP de cuatrimestres anteriores

- Software para segunda entrega:
 - Weka
 - Python
 - R
 - KNIME
 - Orange
 - ...

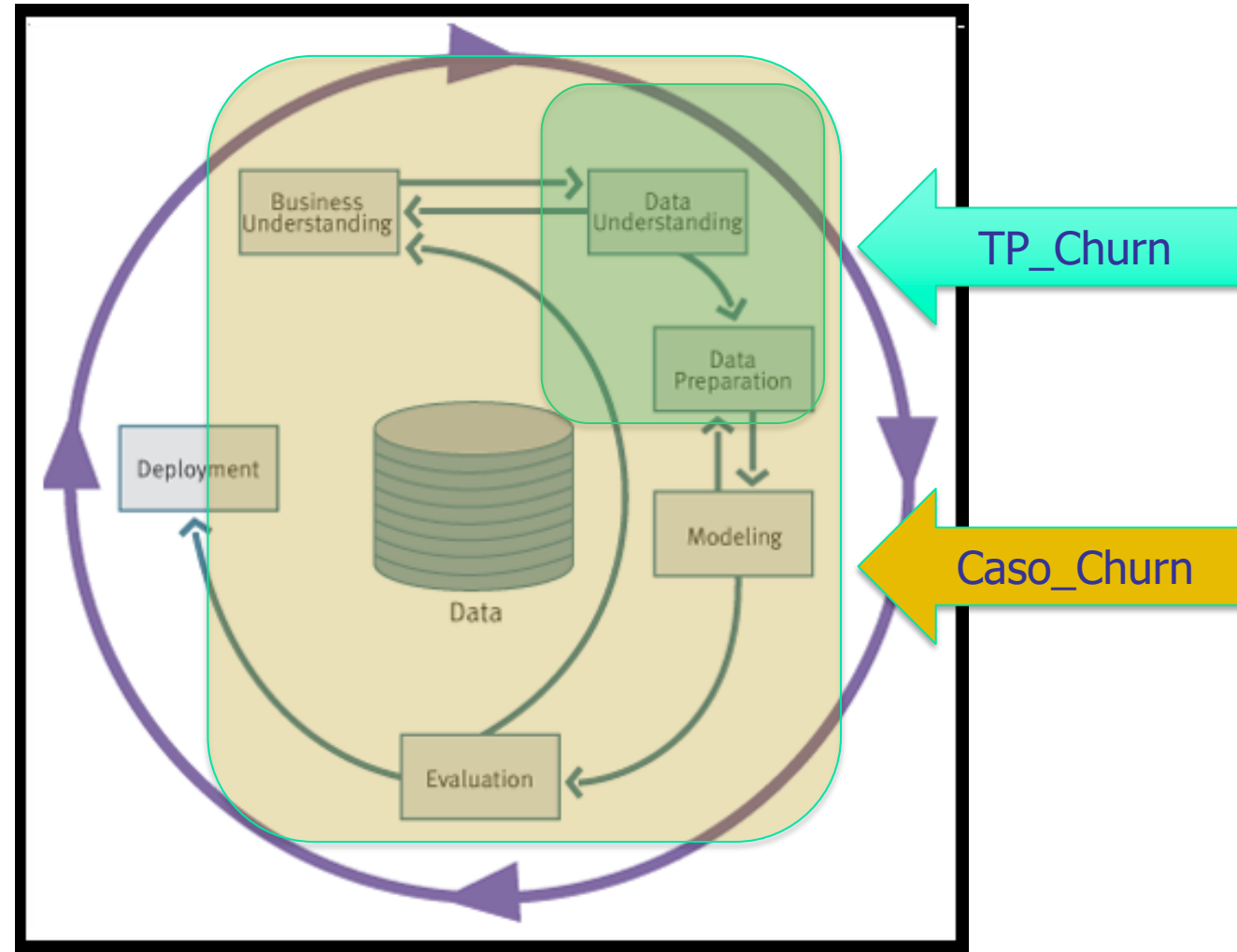
Metodología tradicional para crear modelos



- Identificar el problema y conseguir conocimiento experto.
- Conseguir muchos datos, limpiarlos y completarlos.
- Elegir un método adecuado.
- Entrenar el método con el conjunto de entrenamiento.
- Validar el modelo generado con el conjunto de validación.
- Estimar error.
- Proponer hipótesis.

- KDD
 - From Data Mining to Knowledge Discovery in Databases, Usama Fayyad, Gregory Piatetsky-Shapiro. American Association for Artificial Intelligence, 1996.
<https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>
- SEMMA
 - <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>
- CRISP-DM
 - <https://www.the-modeling-agency.com/crisp-dm.pdf>
- A Parallel Overview:
 - <https://pdfs.semanticscholar.org/7dfe/3bc6035da527deaa72007a27cef94047a7f9.pdf>
- Agile KDD
 - Santana do Nascimento y Oliveira <https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>
- ASD + CRISP-DM
 - Mariscal, Marbán y Segovia: Un enfoque ágil para el desarrollo de proyectos de DM.

Metodologías



ASD SOBRE CRISP-DM

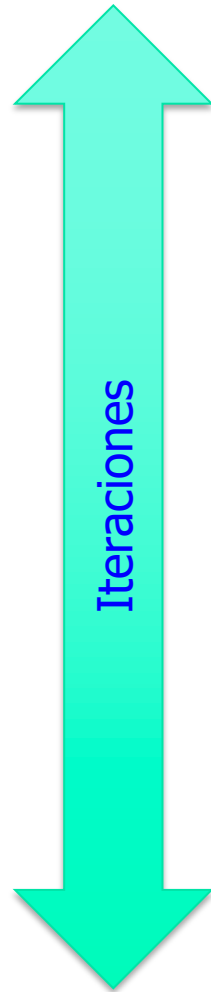
REEMPLAZAR EL CICLO:

PLANIFICACIÓN-DISEÑO-CONSTRUCCIÓN

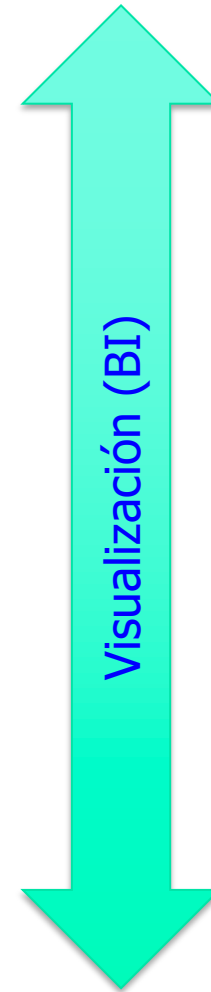
POR:

ESPECULACIÓN-COLABORACIÓN-APRENDIZAJE

Metodología actual para crear modelos



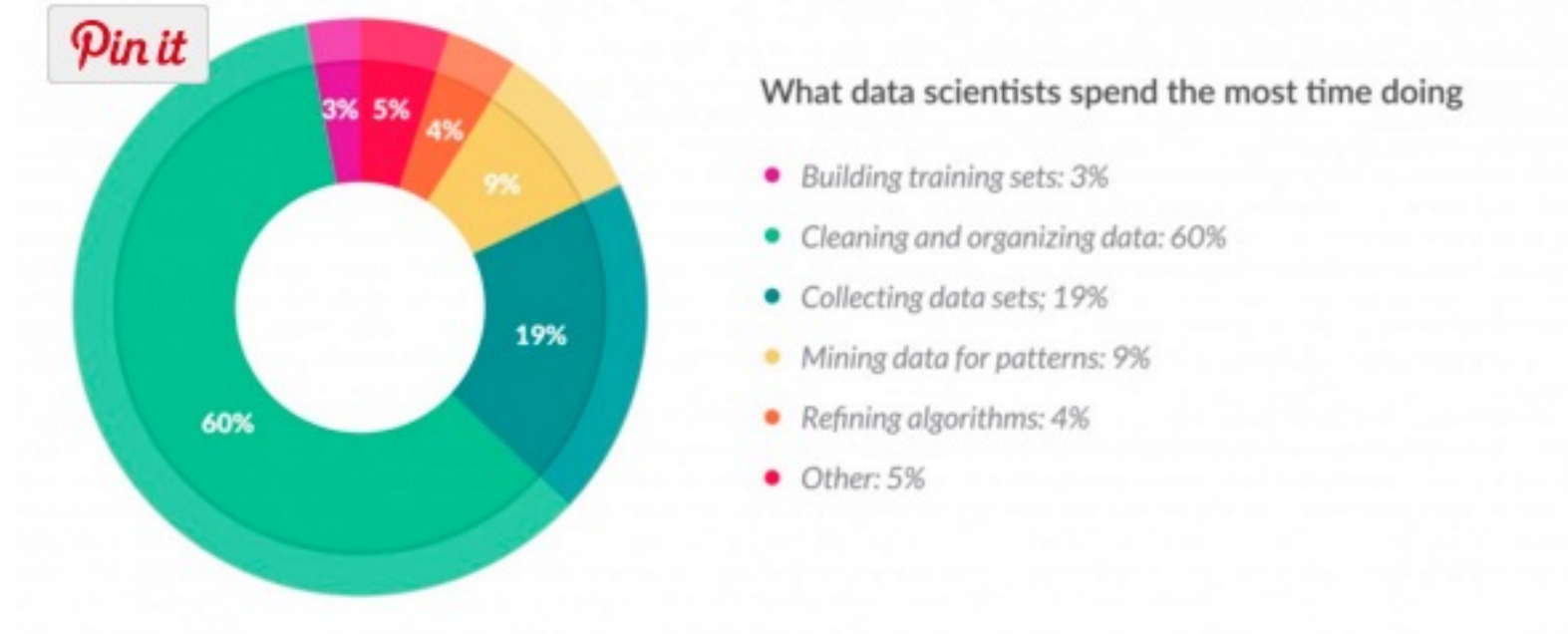
- Identificar el problema y conseguir conocimiento experto.
- Entender la pregunta a responder.
- Comprender los riesgos.
- Entender y preparar los datos.
- Resampling.
- Elegir varios métodos adecuados.
- Desarrollar experimentos con modelos simples.
- Desarrollar experimentos con metaalgoritmos.
- Validar individualmente los modelos simples.
- Ensamblar y validar.
- Visualizar, validar con experto y proponer hipótesis.
- Informar, publicar y testear con externos.
- Volver a empezar.



Metodología actual para crear modelos

- Primera parte
 - Entendimiento del negocio y de los datos
- Segunda parte
 - Modelado, validación y conclusiones

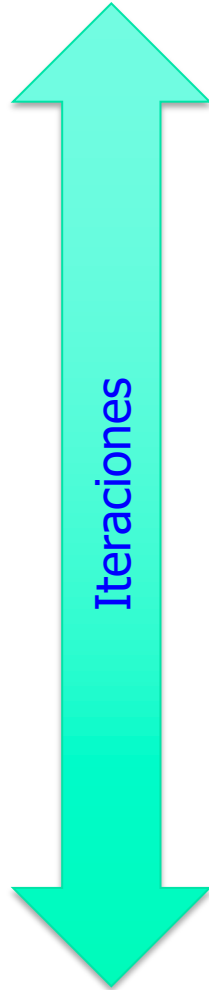
***Data preparation** accounts for about 80% of the work of data scientists*



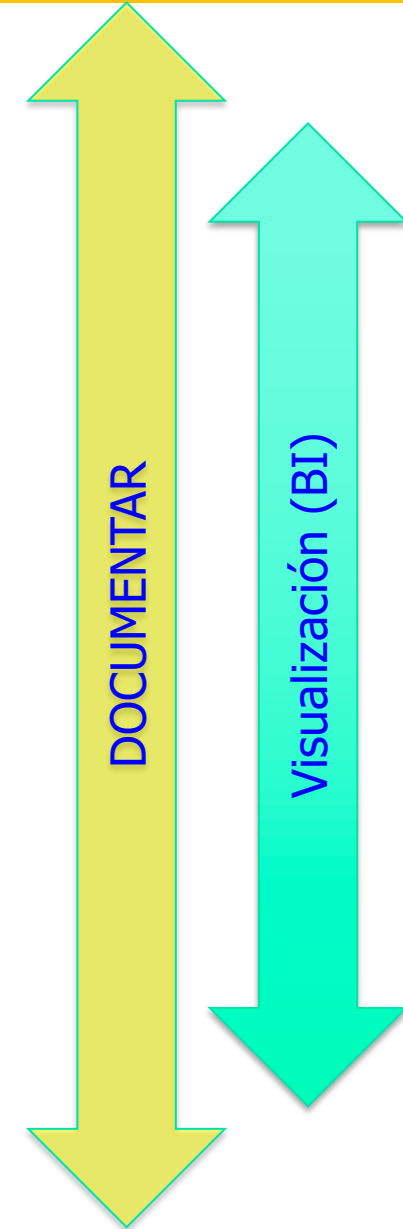
Data scientists spend 60% of their time on cleaning and organizing data. Collecting data sets comes second at 19% of their time, meaning data scientists spend around 80% of their time on preparing and managing data for analysis.

- Obtención
- Ingesta
- Validación, limpieza, compleción
- Preprocesamiento según formatos
- Integración
- Adecuación
 - Outliers
 - Datos faltantes
 - Discretización
 - Normalización / estandarización
 - Derivación / Indicadores
- Documentación
 - Semántica
 - Propagación
 - Linaje
- Curaduría
- Eliminación, síntesis y agregado de variables

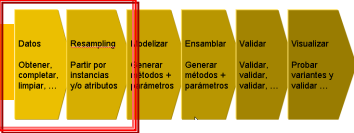
Preparar los datos



- Entender la semántica.
- Entender el ciclo de vida.
- Entender la clase (si la tiene).
- Entender las codificaciones.
- Análisis univariados.
- Estandarizar y/o normalizar.
- Análisis de correlación.
- Aumentar / disminuir volumen.
- Disminuir dimensionalidad.
- Aumentar dimensionalidad:
 - Cocientes, productos, ..., igualdades.
 - Agrupamientos
 - Discretizaciones.
 - Binarizaciones.
 - Tratar outliers y faltantes.
- Si Resampling?
 - Tratar según tipo de dato.

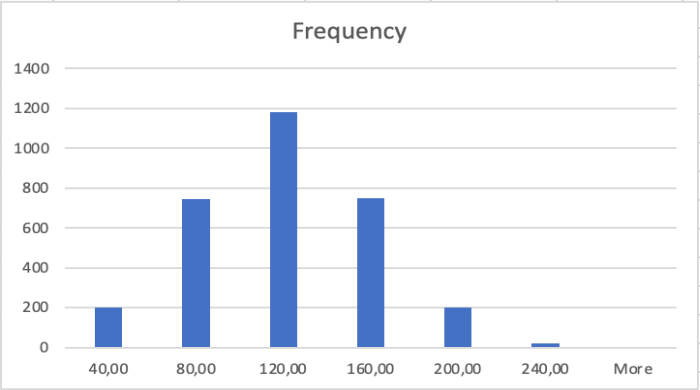


Descripción de datos



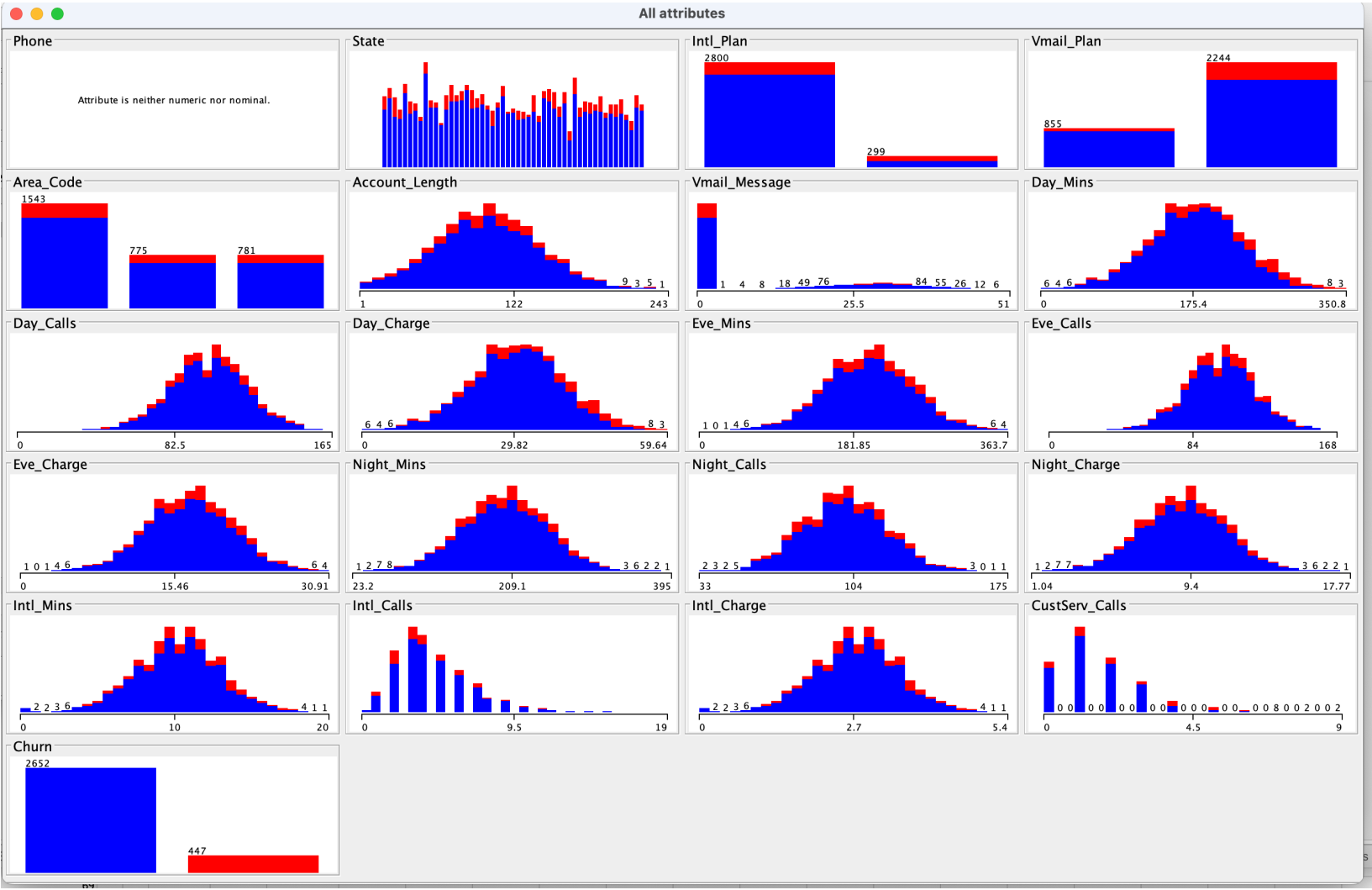
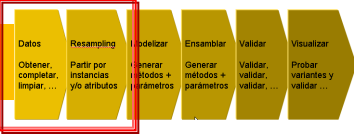
Account_Length		Vmail_Message	
Mean	101,2	Mean	8,1
Standard Error	0,7	Standard Error	0,2
Median	101,0	Median	0,0
Mode	105,0	Mode	0,0
Standard Deviation	39,9	Standard Deviation	13,7
Sample Variance	1588,3	Sample Variance	186,8
Kurtosis	-0,1	Kurtosis	0,0
Skewness	0,1	Skewness	1,3
Range	242,0	Range	51,0
Minimum	1,0	Minimum	0,0
Maximum	243,0	Maximum	51,0
Sum	313569,0	Sum	25023,0
Count	3099,0	Count	3099,0

0	Frequency
40,00	202
80,00	745
120,00	1180
160,00	747
200,00	202
240,00	22
More	1



Row Labels	408	415	510 (blank)	Grand Total
AK	13	23	12	48
AL	23	38	15	76
AR	10	22	15	47
AZ	13	31	13	57
CA	7	16	9	32
CO	25	29	12	66
CT	22	35	9	66
DC	12	26	12	50
DE	11	29	15	55
FL	12	31	20	63
GA	13	21	17	51
HI	14	29	8	51
IA	8	18	15	41
ID	10	37	19	66
IL	14	26	14	54
IN	13	29	20	62
KS	12	33	20	65
KY	15	31	11	57
LA	12	24	10	46
MA	22	26	12	60
MD	15	39	12	66
ME	13	23	21	57
MI	12	37	20	69
MN	19	39	24	82
MO	13	35	11	59
MS	14	31	17	62
MT	17	32	17	66
NC	22	25	13	60
ND	17	26	15	58
NE	13	33	14	60
NH	22	18	11	51
NJ	14	31	19	64
NM	15	32	11	58
NV	14	34	17	65
NY	18	42	15	75
OH	20	36	16	72
OK	15	24	13	52
OR	14	40	17	71
PA	14	19	9	42
RI	11	33	16	60
SC	10	26	16	52
SD	14	25	15	54
TN	10	27	11	48
TX	20	36	15	71
UT	10	35	23	68
VA	25	34	17	76
VT	16	34	19	69
WA	23	24	12	59
WI	20	31	19	70
WV	18	48	30	96
WY	16	40	18	74
(blank)				
Grand Total	775	1543	781	3099

Datos vs clase



Correlaciones



	<i>Day_Mins</i>	<i>Day_Calls</i>	<i>Day_Charge</i>	<i>Eve_Mins</i>	<i>Eve_Calls</i>	<i>Eve_Charge</i>	<i>Night_Mins</i>	<i>Night_Calls</i>	<i>Night_Charge</i>
Day_Mins	1,000								
Day_Calls	0,012	1,000							
Day_Charge	1,000	0,012	1,000						
Eve_Mins	0,002	-0,020	0,002	1,000					
Eve_Calls	0,031	0,007	0,031	-0,008	1,000				
Eve_Charge	0,002	-0,020	0,002	1,000	-0,008	1,000			
Night_Mins	0,002	0,017	0,002	-0,014	0,001	-0,014	1,000		
Night_Calls	0,030	-0,008	0,030	0,013	0,015	0,013	0,008	1,000	
Night_Charg	0,002	0,017	0,002	-0,014	0,001	-0,014	1,000	0,008	1,000

Metodología actual para crear modelos

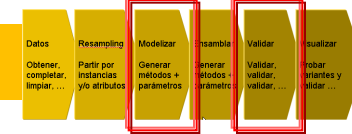
- Primera parte
 - Entendimiento del negocio y de los datos
- Segunda parte
 - Modelado, validación y conclusiones

■ Modelado

- Sólo árboles

■ Validación

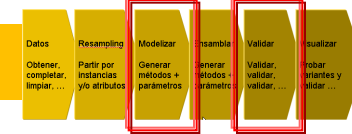
- Indicadores
- Sobre training?
- Sobre testing?
- CV?



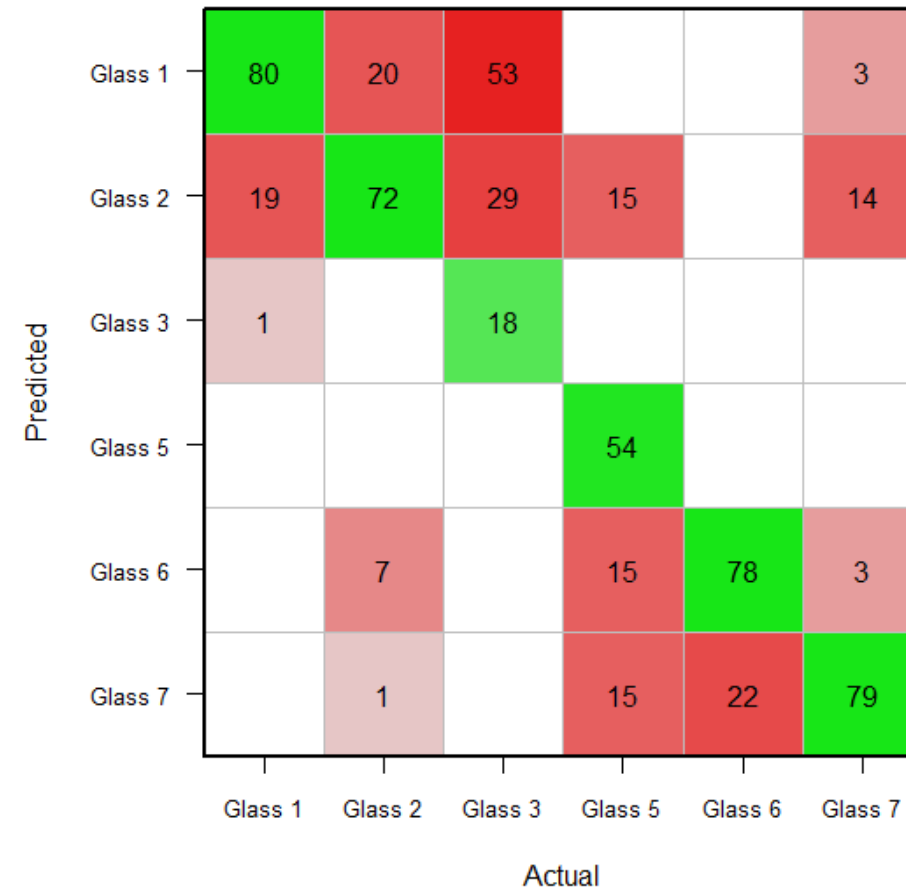
Medidas de Performance

- Un modelo tiene una **exactitud (*accuracy*)** del 95%.
 - O sea, de cada 100 instancias, clasifica bien 95.
- ¿Qué significa esto?
- Según la distribución de clases en el dominio, 95% puede ser muy bueno o pésimo.
- No dice nada sobre el **tipo de aciertos y errores** que comete el modelo.
- Veamos otras medidas de performance más útiles...

Matriz de confusión



	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative





=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2915
Incorrectly Classified Instances	184
Kappa statistic	0.7391
Mean absolute error	0.0949
Root mean squared error	0.2301
Relative absolute error	38.4255 %
Root relative squared error	65.4813 %
Total Number of Instances	3099

94.0626 %
5.9374 %

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.981	0.300	0.951	0.981	0.966	0.744	0.870	0.952	False.
	0.700	0.019	0.862	0.700	0.773	0.744	0.870	0.768	True.
Weighted Avg.	0.941	0.259	0.938	0.941	0.938	0.744	0.870	0.926	

=== Confusion Matrix ===

a	b	<-- classified as
2602	50	a = False.
134	313	b = True.

Matriz de Confusión:
(Clasificación binaria)

tp: true positives
tn: true negatives
fp: false positives
fn: false negatives

	SPAM (predicho)	NO SPAM (predicho)
SPAM (real)	2739 tp	56 fn
NO SPAM (real)	4 fp	1042 tn

Precisión y Recall (“exhaustividad”):

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

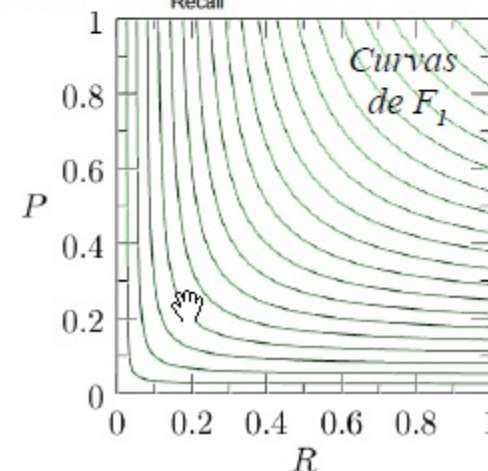
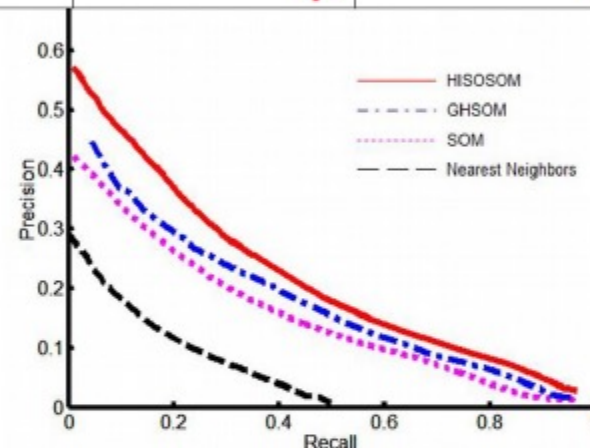
$$F\text{-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Media armónica. También llamada F_1 score.

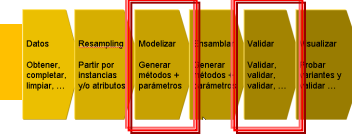
Fórmula general:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

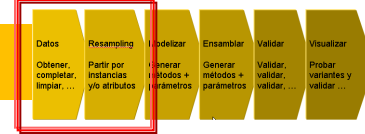
F_2 enfatiza recall; $F_{0.5}$ enfatiza precision.



Entender los riesgos



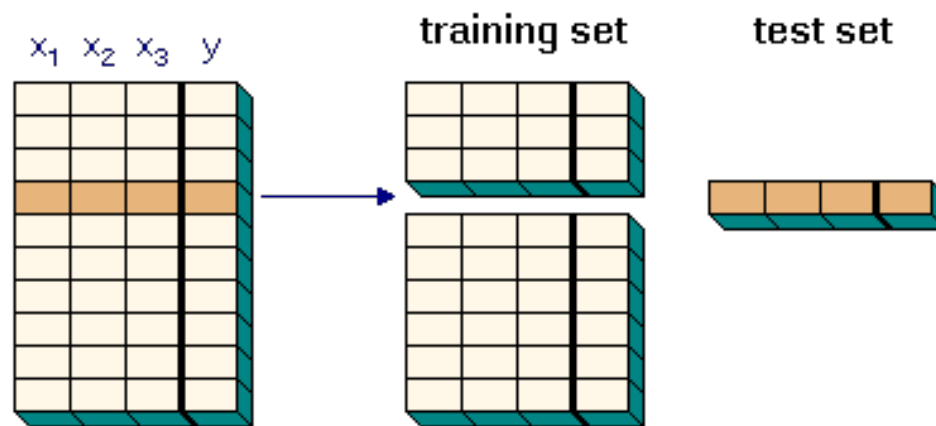
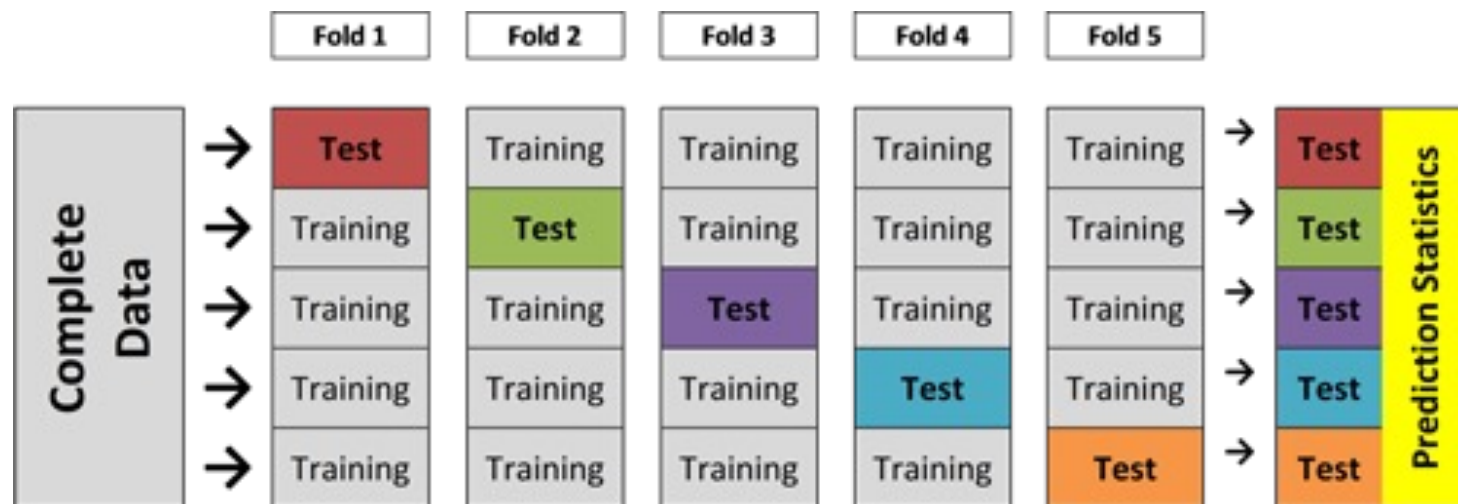
		PROBABILIDAD				
		Raro	Poco probable	Posible	Muy probable	Casi seguro
CONSECUENCIAS	Despreciable	Bajo	Bajo	Bajo	Medio	Medio
	Menores	Bajo	Bajo	Medio	Medio	Medio
	Moderadas	Medio	Medio	Medio	Alto	Alto
	Mayores	Medio	Medio	Alto	Alto	Muy alto
	Catastróficas	Medio	Alto	Alto	Muy alto	Muy alto



■ Pautas metodológicas:

- Usar el train para elegir las variables.
- Usar validación independiente para determinar la cantidad de variables a retener.
- Hacer una selección final con train+test.
- Estimar el error con el test set.

Cross Validation



Validación Cruzada

- ¿Qué puede pasar si tenemos mala suerte al separar los datos para entrenamiento/validación?
- k -Fold Cross Validation:
 - 1) Desordenar los datos.
 - 2) Separar en k folds del mismo tamaño.
 - 3) Para $i = 1 \dots k$:
 - Entrenar sobre todos los folds menos el i .
 - Evaluar sobre el fold i .

- Ej. para $k=5$:

Entrenamiento
Validación

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	→ Resultado 1
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	→ Resultado 2
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	→ Resultado 3
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	→ Resultado 4
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	→ Resultado 5

→ Promedio

Datos de Test

- Lo antes posible, hay que separar un conjunto de **datos de test** (*test set*), y **NO TOCARLOS** hasta el final.
- Todas las pruebas y ajustes se hacen sobre el conjunto de **datos de desarrollo** (*dev set*).
- Cuando termina el desarrollo, se evalúa sobre los datos de test separados. La estimación de performance será más **realista**.
- ¡No volver atrás!

DESARROLLO
(Elección de algoritmos, cross-validation, etc.)

TEST





- Cuántos data sets?:
 - $\text{DataSet} = \text{DevSet} + \text{TestSet} (*2?)$.
 - $\text{DevSet} = \text{TrainSet} + \text{ValidationSet} (\text{CV}?)$.
 - Usar el train para elegir las variables.
 - Usar validación independiente para determinar la cantidad de variables a retener.
 - Hacer una selección final con DevSet.
 - Desarrollar con DevSet (CV).
 - Estimar el error con el TestSet.

DOCUMENTAR

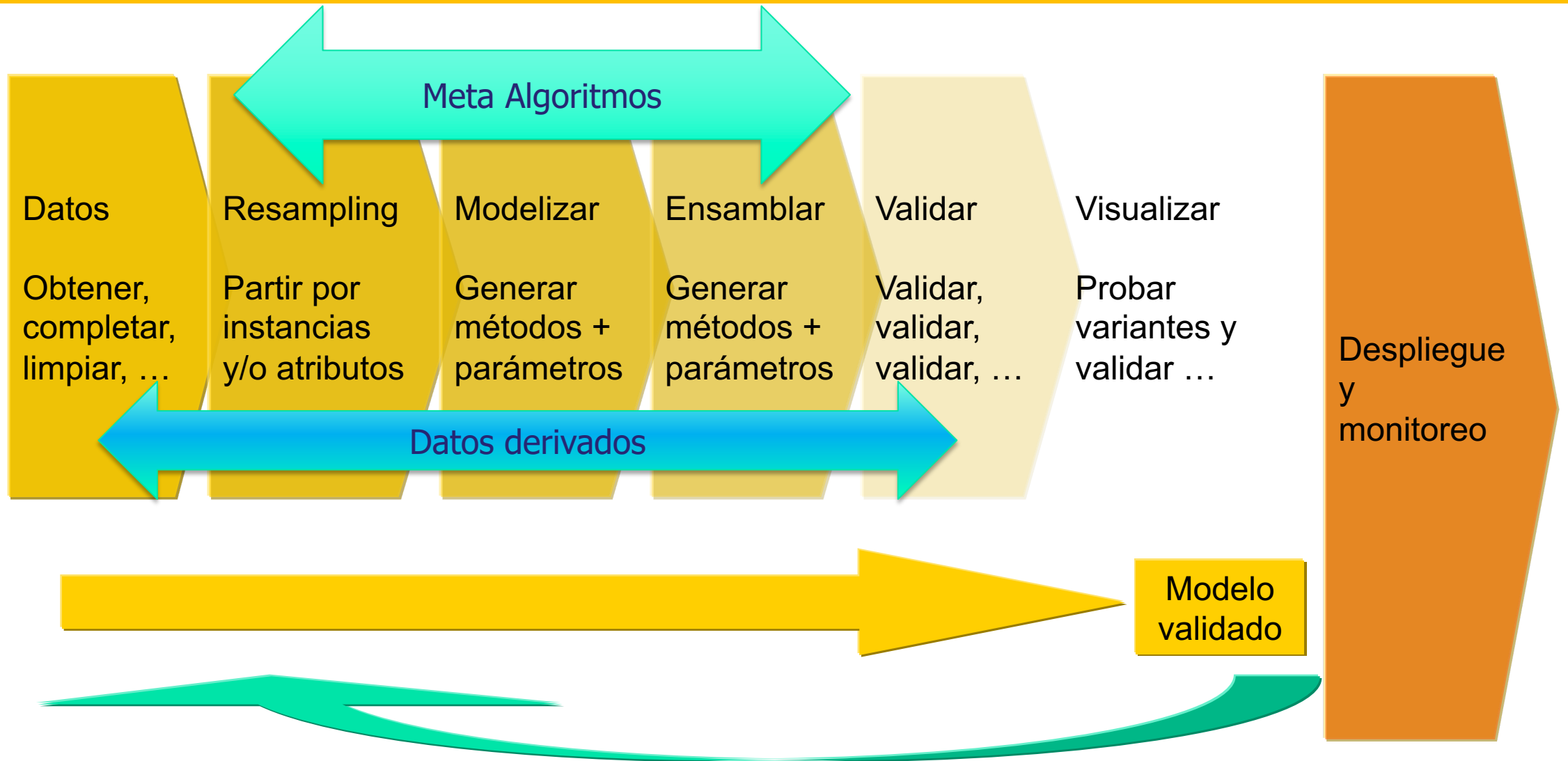
- Metodologías para DM
- Preparación de datos
- Resampling
- Modelización
- Ensamble
- Validación
- Publicación

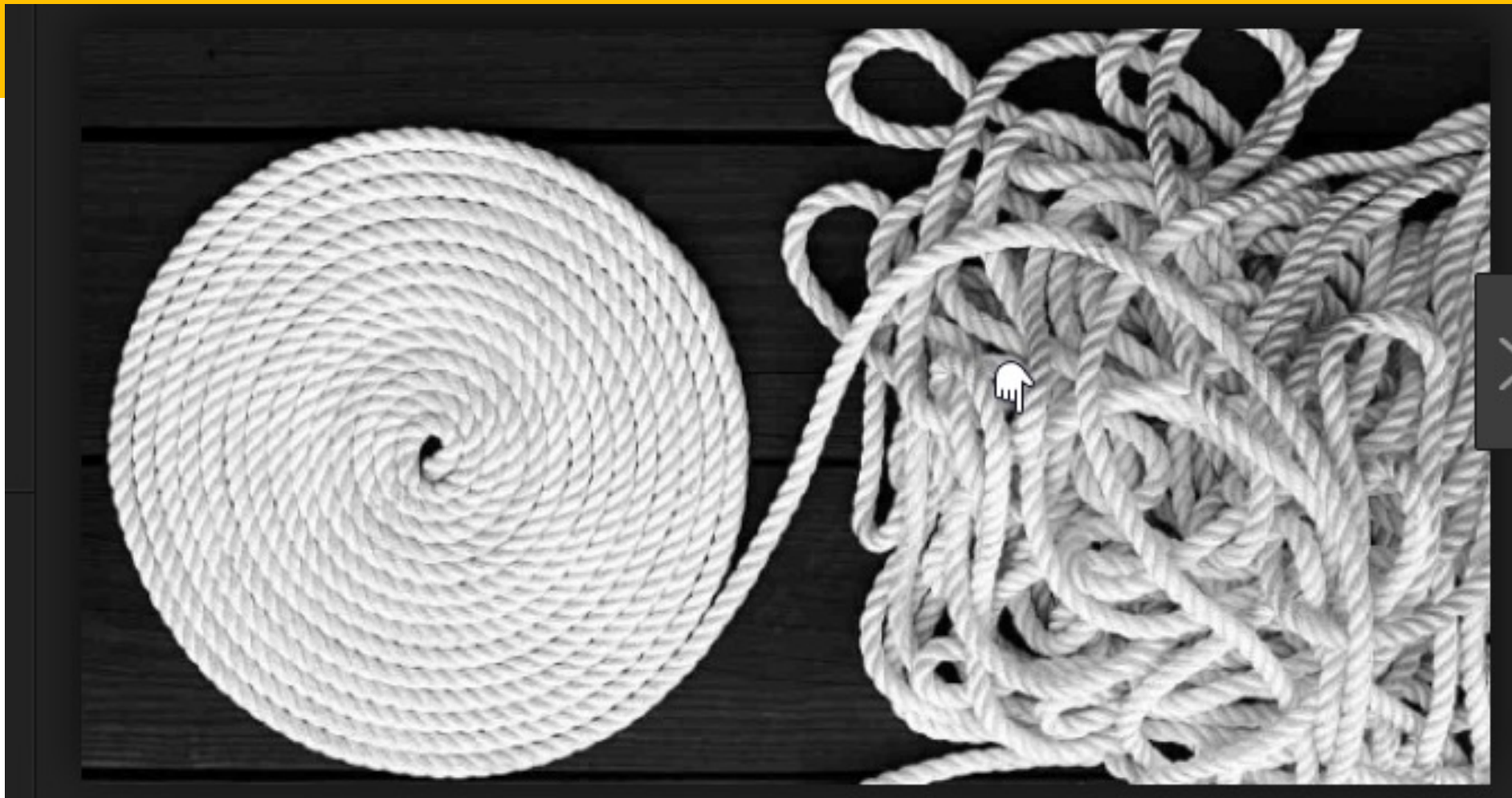


Asegurar pruebas de integridad

- **Detalle:**
 - Código
 - Sistema de ticketing
 - Pruebas
- **Presentación:**
 - Visualización
 - Informe gerencial
 - Informe interno de cierre
- **Issues:**
 - Mejoras al reservorio (data y metadata).
 - Lecciones aprendidas.
 - Exposiciones internas.

Metodología actual para crear modelos





La confusión está clarísima !