PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

UC | Chile

UC | Chile

Aprendizaje Supervisado: Overfitting y underfitting

Contenidos

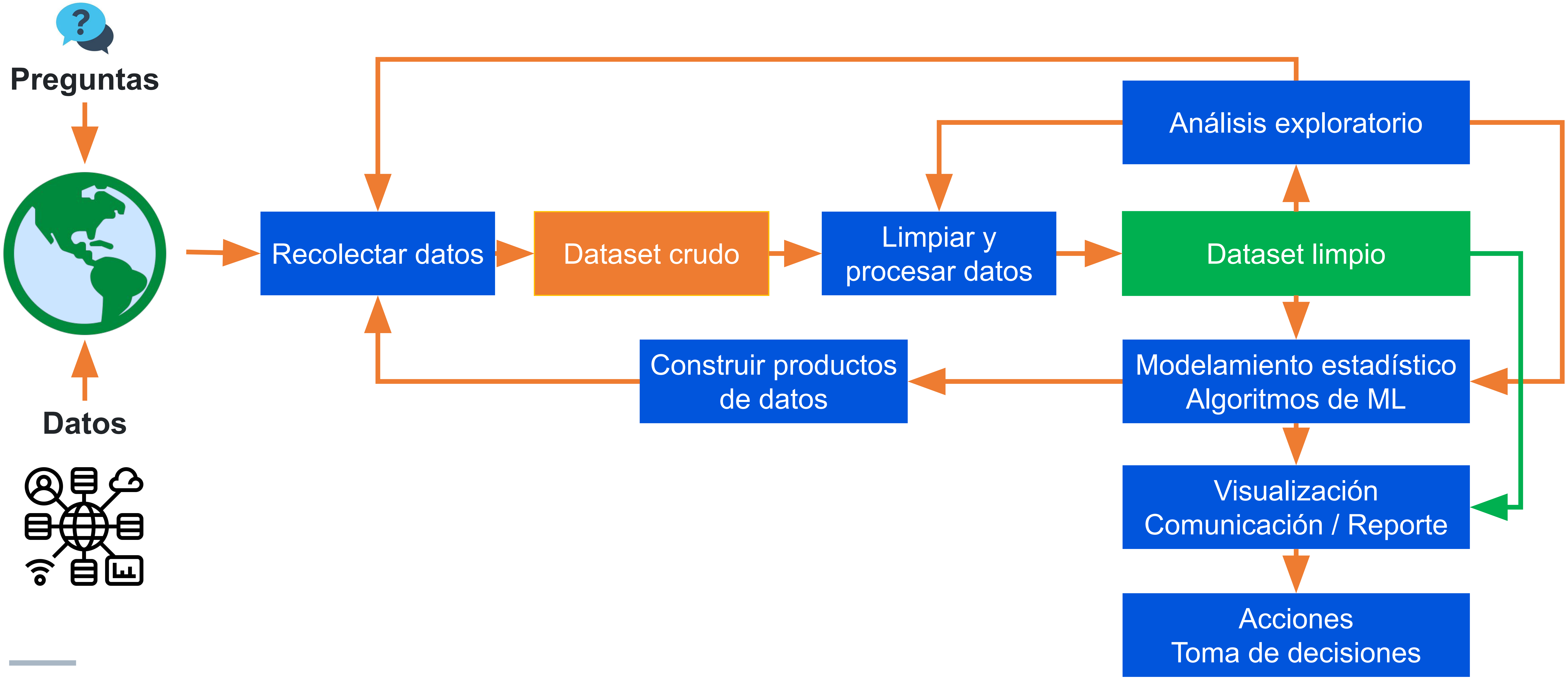
Tema 1

Repaso: algoritmos de aprendizaje supervisado

Tema 2

Overfitting y underfitting

Proceso de Ciencia de Datos



Fuente: Adaptado de O'Neil, Cathy,Schutt, Rachel. "Doing Data Science", O'Reilly Media.

ALGORITMOS DE APRENDIZAJE SUPERVISADO

ALGORITMOS DE APRENDIZAJE DE MÁQUINA

Algoritmos de aprendizaje de máquina (ML) métodos computacionales que utilizan data anterior (i.e. experiencia) para generar modelos o programas capaces de realizar tareas como predecir, clasificar, agrupar, ordenar o reducir dimensionalidad.

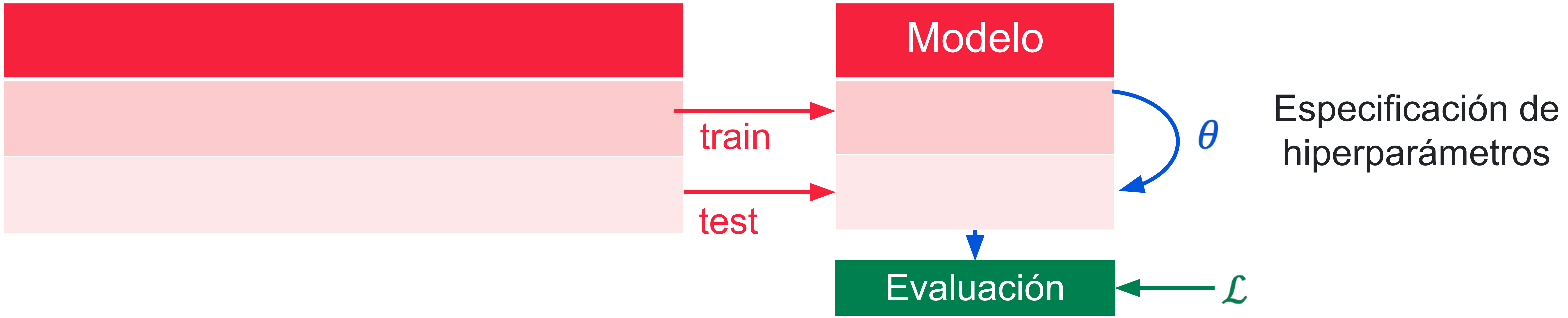
- Para una tarea dada, pueden proponerse **múltiples algoritmos** posibles.
- El éxito de un algoritmo de ML se evalúa en base a **métricas** de precisión, eficiencia y tiempo computacional.
- La elección del algoritmo a usar dependerá de: **contexto y complejidad del problema**, suposiciones de base, tamaño y variedad de la **data** disponible.
- Implementación.

APRENDIZAJE SUPERVISADO

El objetivo es realizar predicciones precisas para ***nuevos datos*** con características similares a los datos usados para construir el modelo → generalización

Entrenamiento y testeo:

Hiperparámetros (θ) ➔ parámetros libres del modelo que no son determinados por el algoritmo, sino entregados como input



MODELOS DE REGRESIÓN

- En un problema de regresión, buscamos predecir el valor de una variable a partir del valor de otras variables.

Ejemplo: Predecir el consumo de combustible de un auto a partir de sus características de diseño.

n observaciones

i=1,2,...,n

	car_name	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

$Y = y_1, \dots, y_n$

outcome / variable dependiente respuesta

$X = X_1, \dots, X_p$

X_j

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

predictores /variable independiente/ features

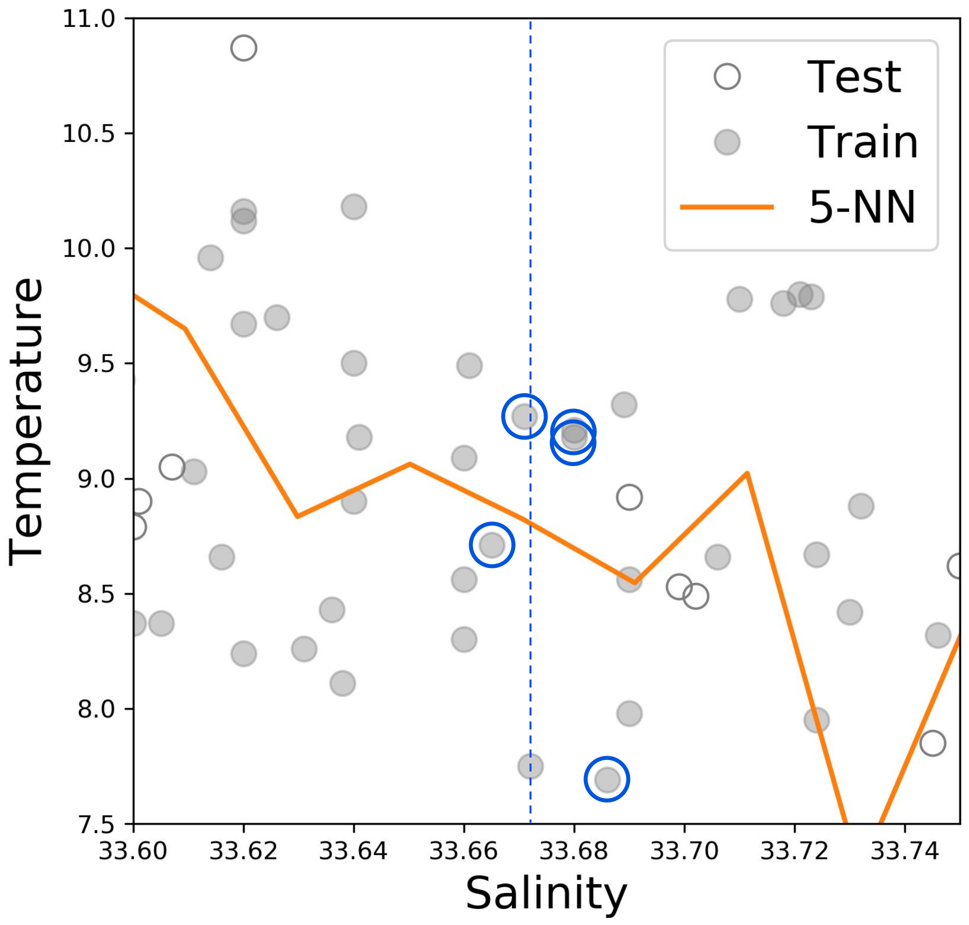
MODELOS DE REGRESIÓN: kNN

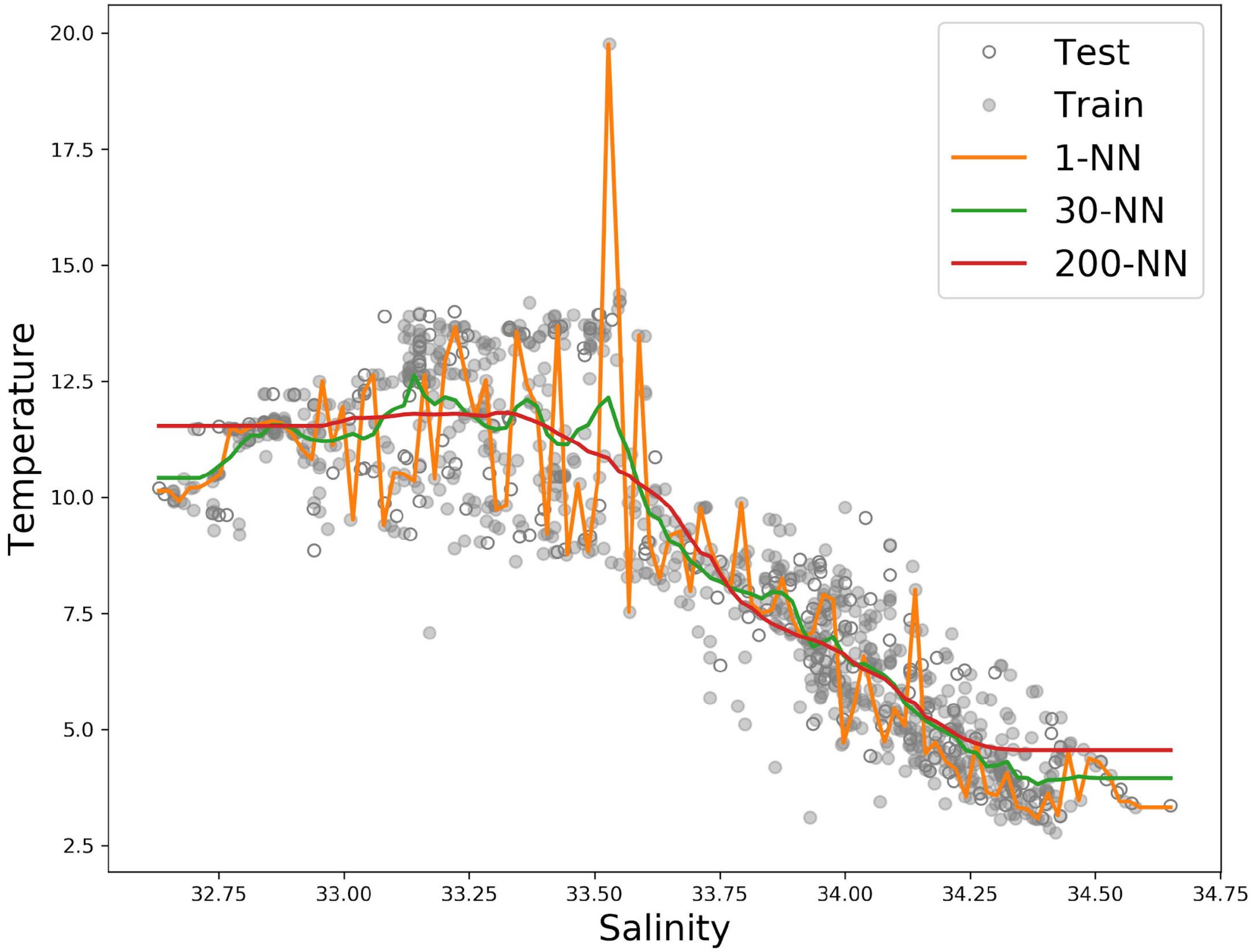
Regresión kNN

$$\hat{y}_i = \frac{1}{k}\sum_{j=1}^{k} y_{i_j}$$

Donde $\{x_{i1}, \ldots, x_{ik}\}$ son las **k** observaciones más similares (cercanas) a x_i

- Requiere normalización de variables.





MODELOS DE REGRESIÓN: REGRESIÓN LINEAL Y MULTILINEAL

Regresión Lineal

Y depende de una variable predictora.

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$$

Regresión Multilineal

Y depende de varias variables predictoras.

$$Y = f\left(X_1, \dots, X_J\right) + \epsilon = \beta_0 + \beta_1 X_1 \dots + \beta_J X_J + \epsilon$$

$$\Rightarrow \boldsymbol{Y = \beta X}$$

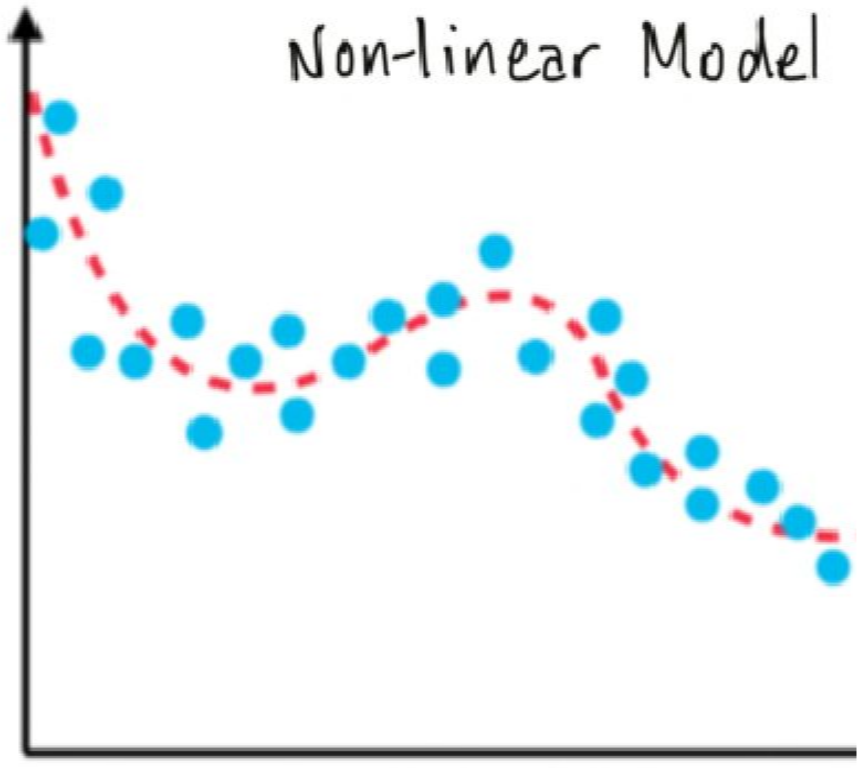
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

$$\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 X_1 \dots + \beta_J X_J)\right)^2 \rightarrow \hat{\beta} = \operatorname{argmin} \mathcal{L}(\beta)$$

MODELOS DE REGRESIÓN: REGRESIÓN POLINOMIAL

Regresión polinomial:

$$Y = f_\beta(X)$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$



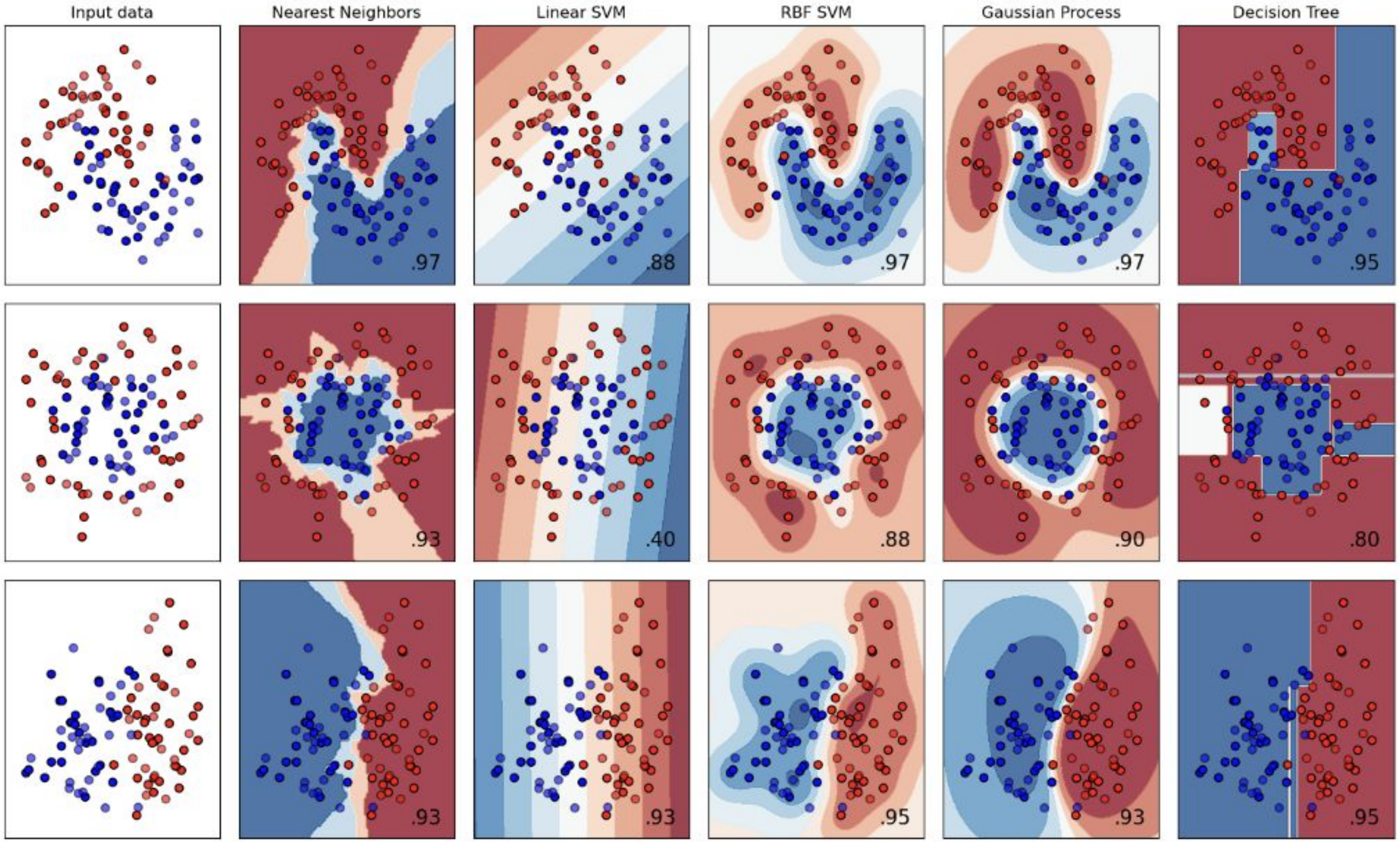
f : una función no-lineal

β : vector de parámetros de f

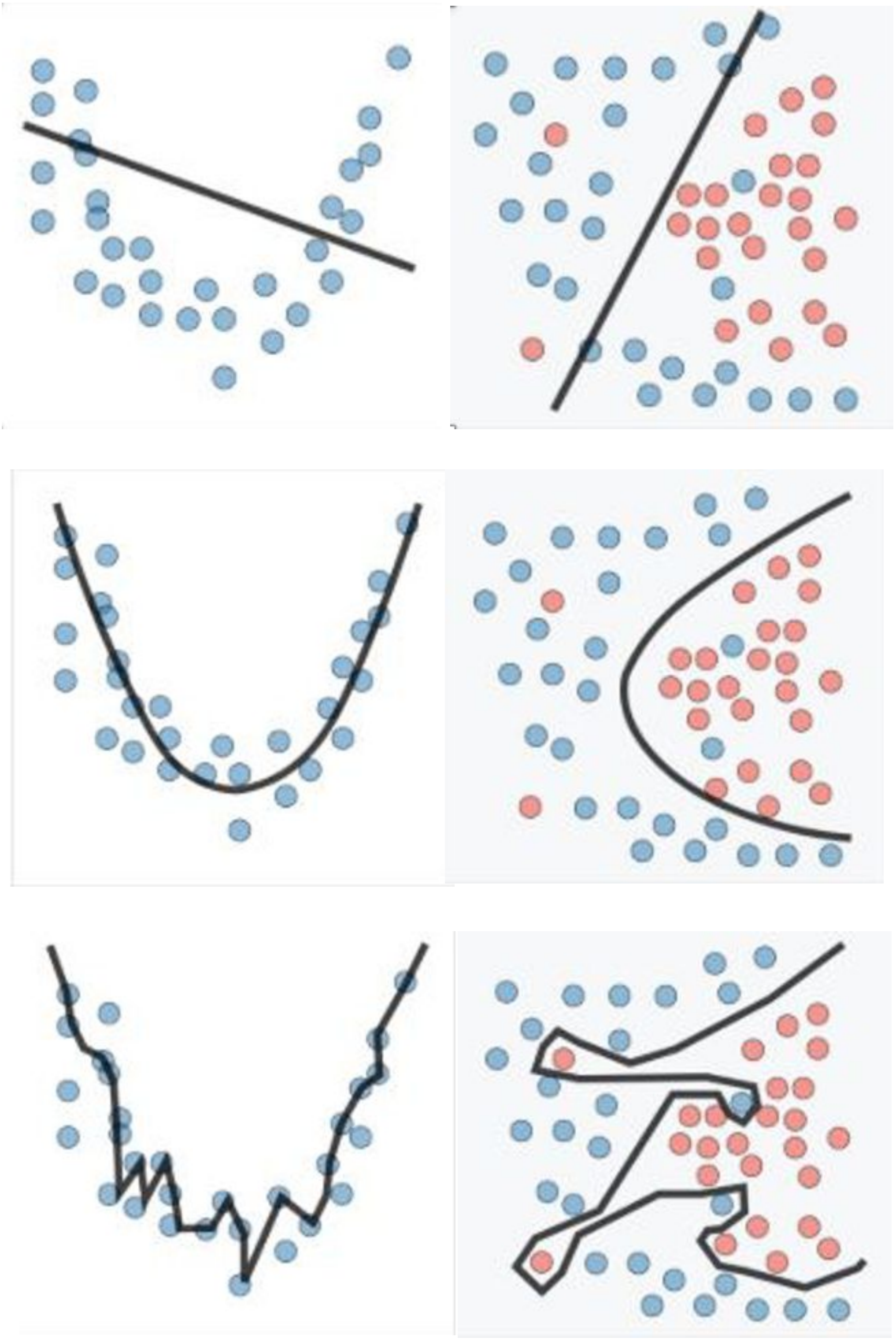
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^M \\ 1 & x_2^1 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

$$\Rightarrow \boldsymbol{Y} = \boldsymbol{\beta X}$$

MODELOS DE CLASIFICACIÓN



APRENDIZAJE SUPERVISADO: OVERFITTING



Regresión Clasificación

Underfitting (subajuste)
Alto error de entrenamiento
Error de prueba similar a error de entrenamiento

Óptimo
Error de entrenamiento levemente más bajos que error de prueba

Overfitting (sobreajuste)
Muy bajo error de entrenamiento
Error de prueba mucho mayor a error de entrenamiento

APRENDIZAJE SUPERVISADO: OVERFITTING

Si el modelo se ajusta muy cercanamente a la data de entrenamiento, pero falla al generalizar o predecir la data de prueba □ overfitting

Factores que influyen en overfitting:

1. Complejidad del modelo (*d*)

- Demasiado simple □ underfitting
- Demasiado complejo □ overfitting

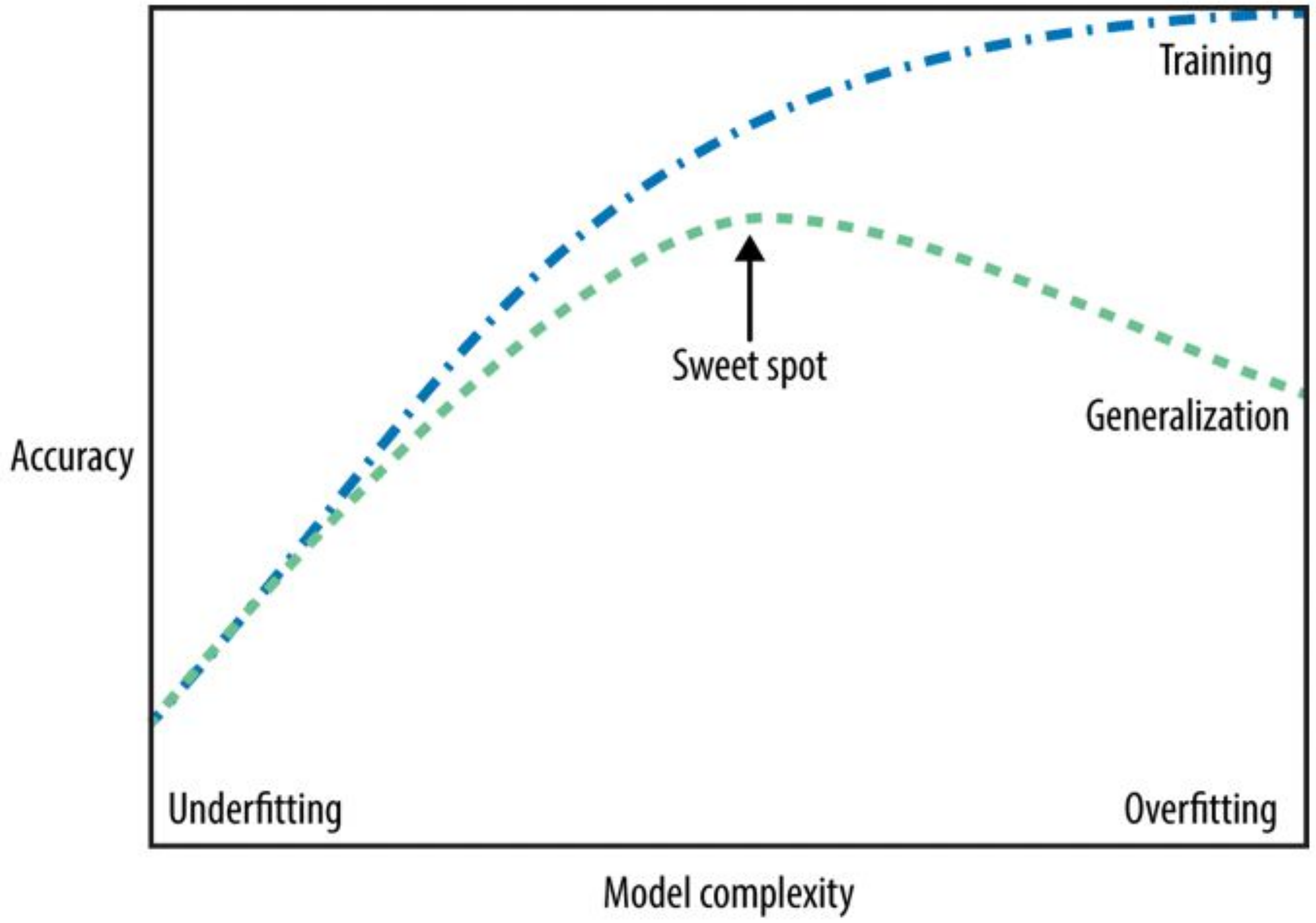


Figure 2-1. Trade-off of model complexity against training and test accuracy

APRENDIZAJE SUPERVISADO: OVERFITTING

Factores que influyen en overfitting:

2. Nº de datos de entrenamiento (*N*)

- A mayor cantidad y variedad de datos, más complejo puede ser el modelo sin caer en overfitting

3. Magnitud del ruido

- Mientras más ruidosos son los datos, mayor posibilidad de sobreajuste.

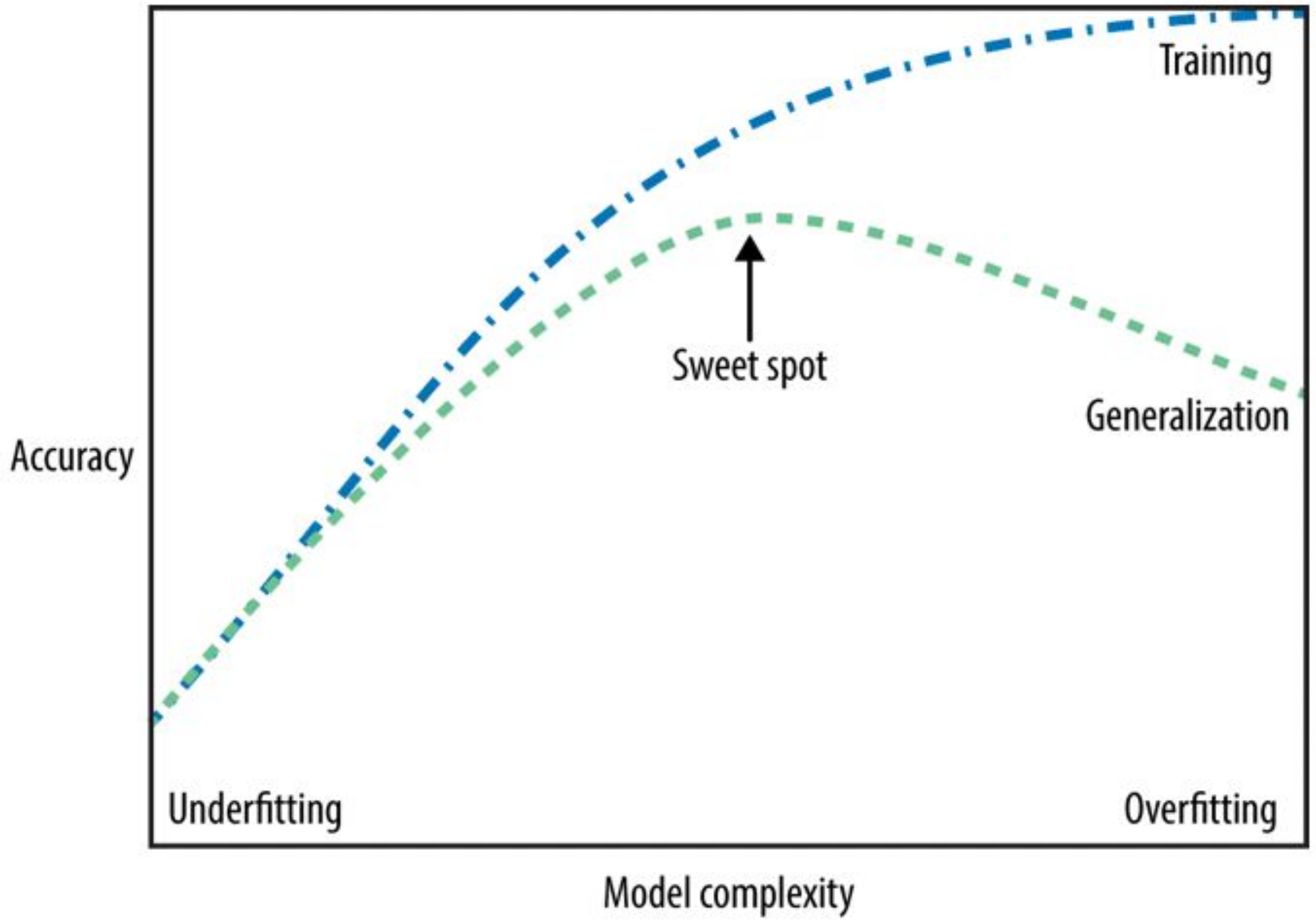
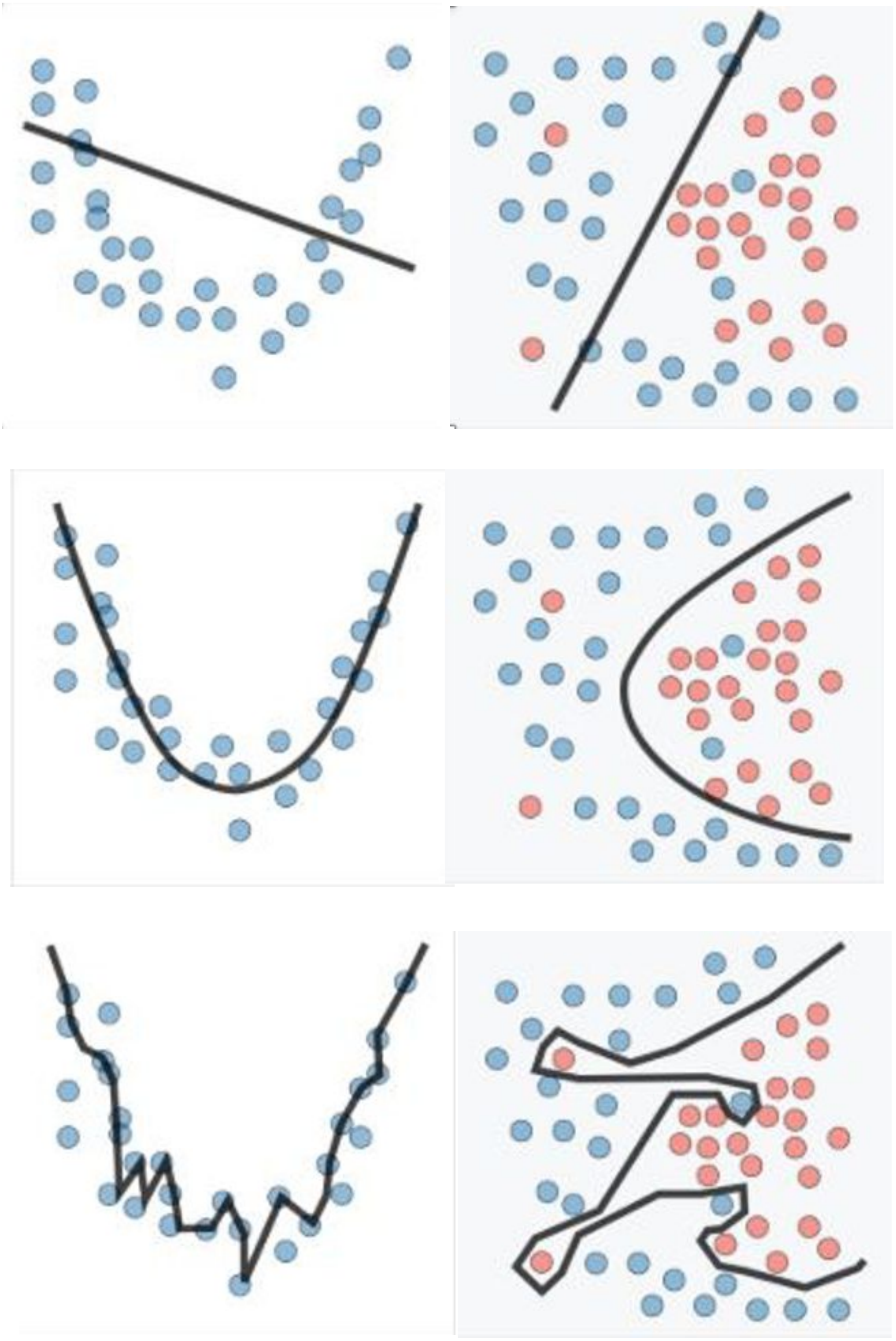


Figure 2-1. Trade-off of model complexity against training and test accuracy

APRENDIZAJE SUPERVISADO: OVERFITTING



Regresión Clasificación

Underfitting (subajuste)
Alto error de entrenamiento
Error de prueba similar a error de entrenamiento

Complejizar modelo
Agregar features

Óptimo
Error de entrenamiento levemente más bajos que error de prueba

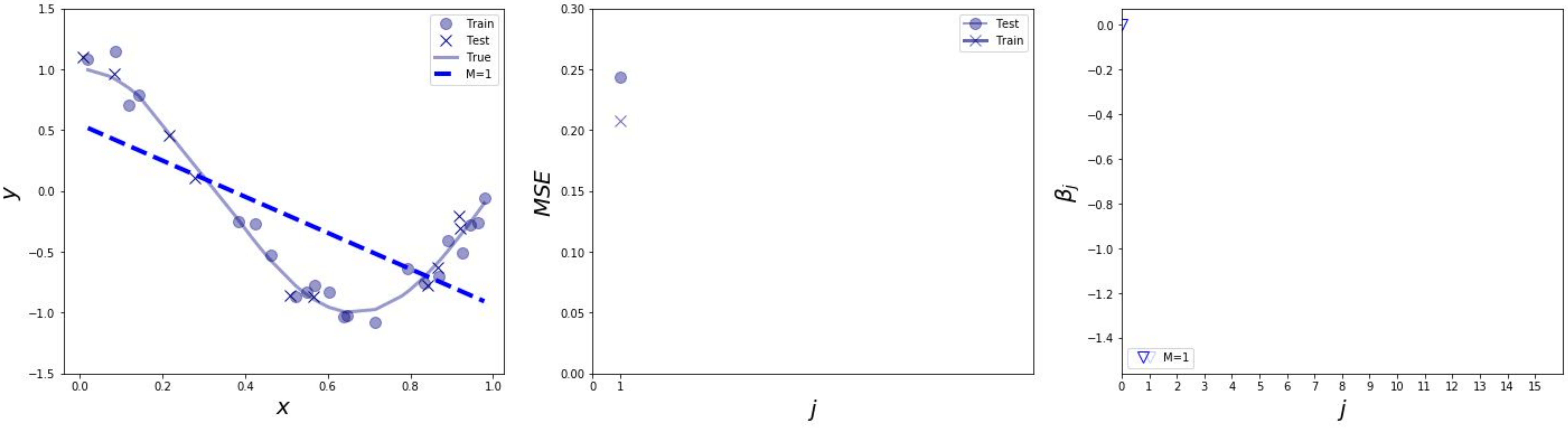
Overfitting (sobreajuste)
Muy bajo error de entrenamiento
Error de prueba mucho mayor a error de entrenamiento

Simplificar modelo
Buscar más datos
Regularización

REGRESIÓN Y OVERFITTING

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Supongamos un modelo de regresión polinomial para un conjunto de **n=30** datos

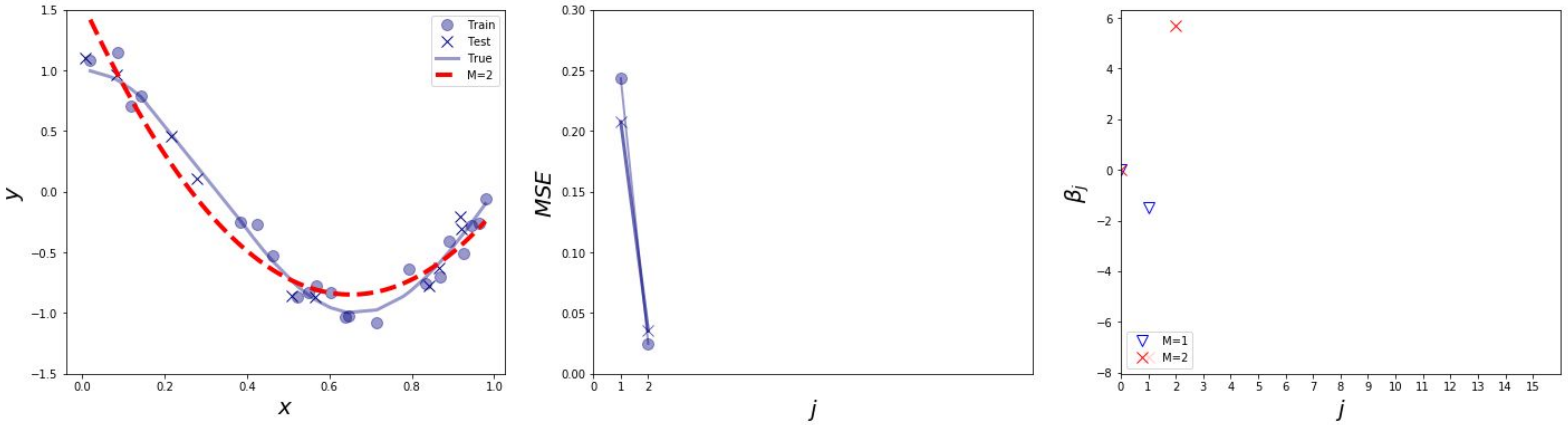


Ajuste lineal □ underfitting

REGRESIÓN Y OVERFITTING

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Supongamos un modelo de regresión polinomial para un conjunto de **n=30** datos

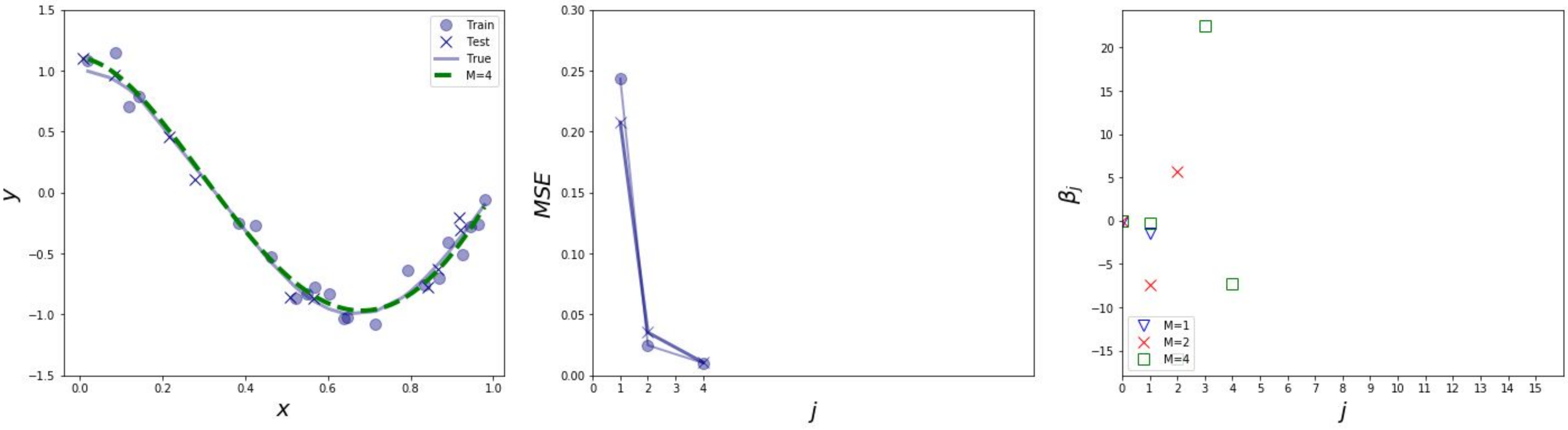


Disminuye el error ☺

REGRESIÓN Y OVERFITTING

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Supongamos un modelo de regresión polinomial para un conjunto de **n=30** datos

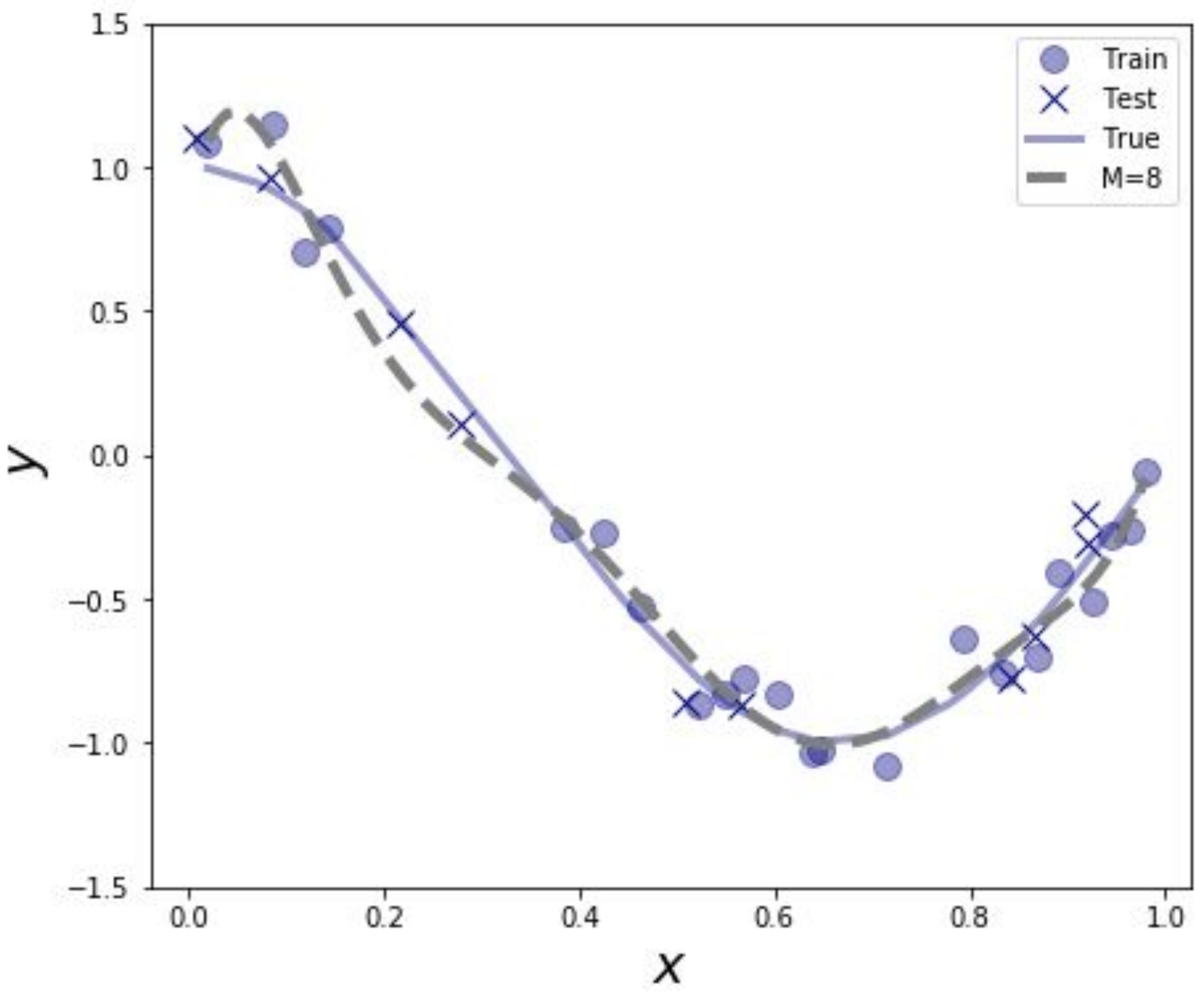


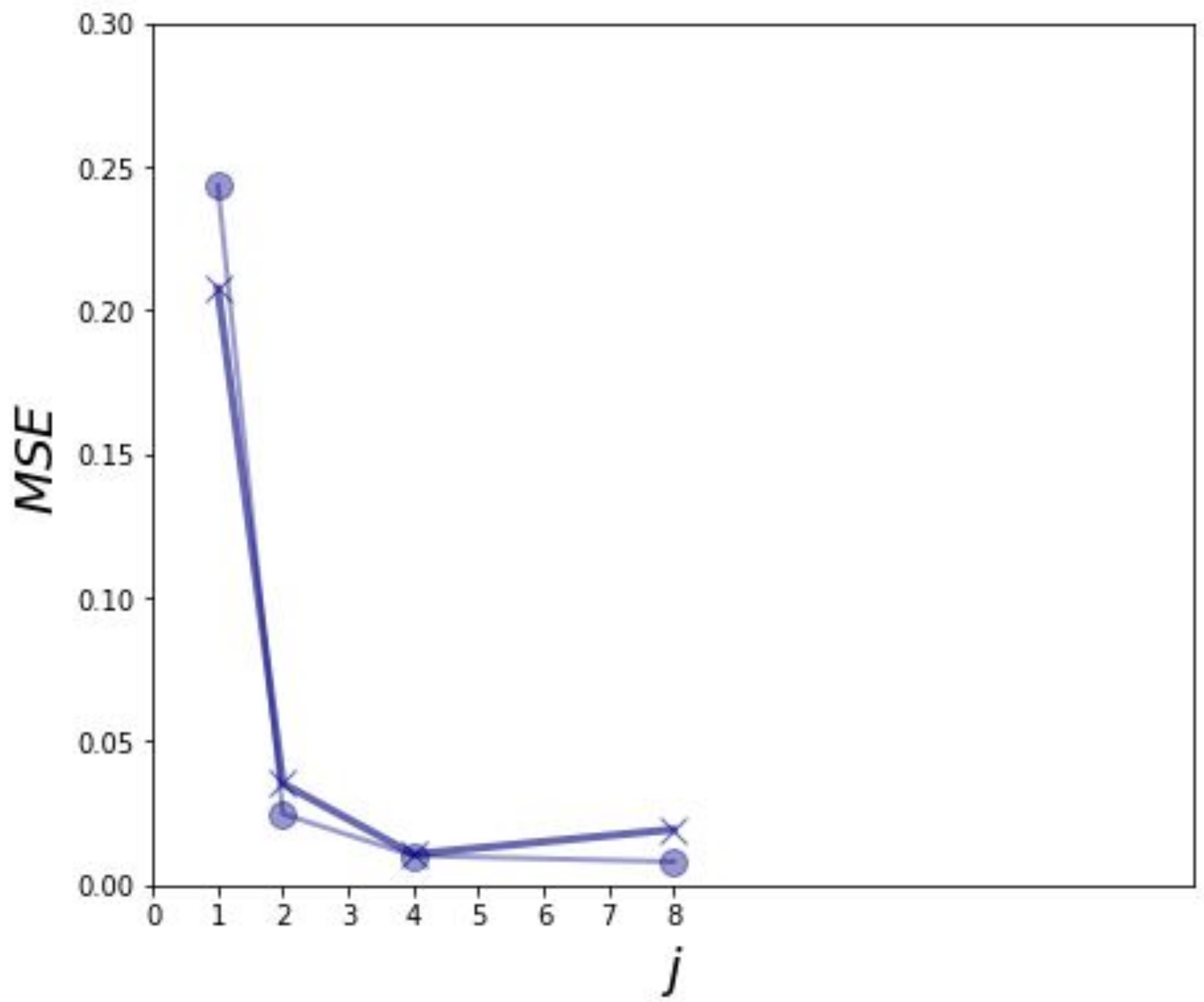
Disminuye el error ☺

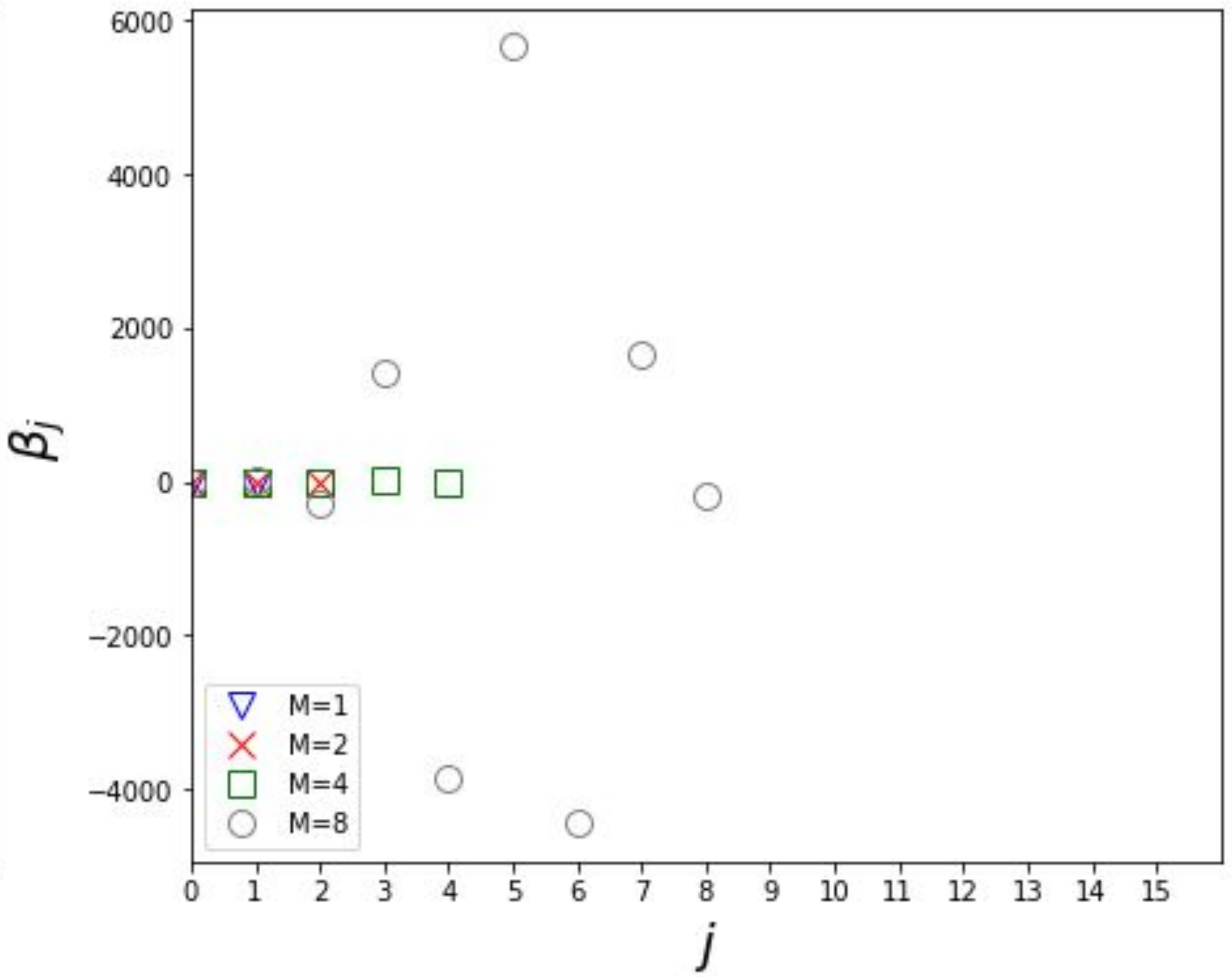
REGRESIÓN Y OVERFITTING

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Supongamos un modelo de regresión polinomial para un conjunto de **n=30** datos







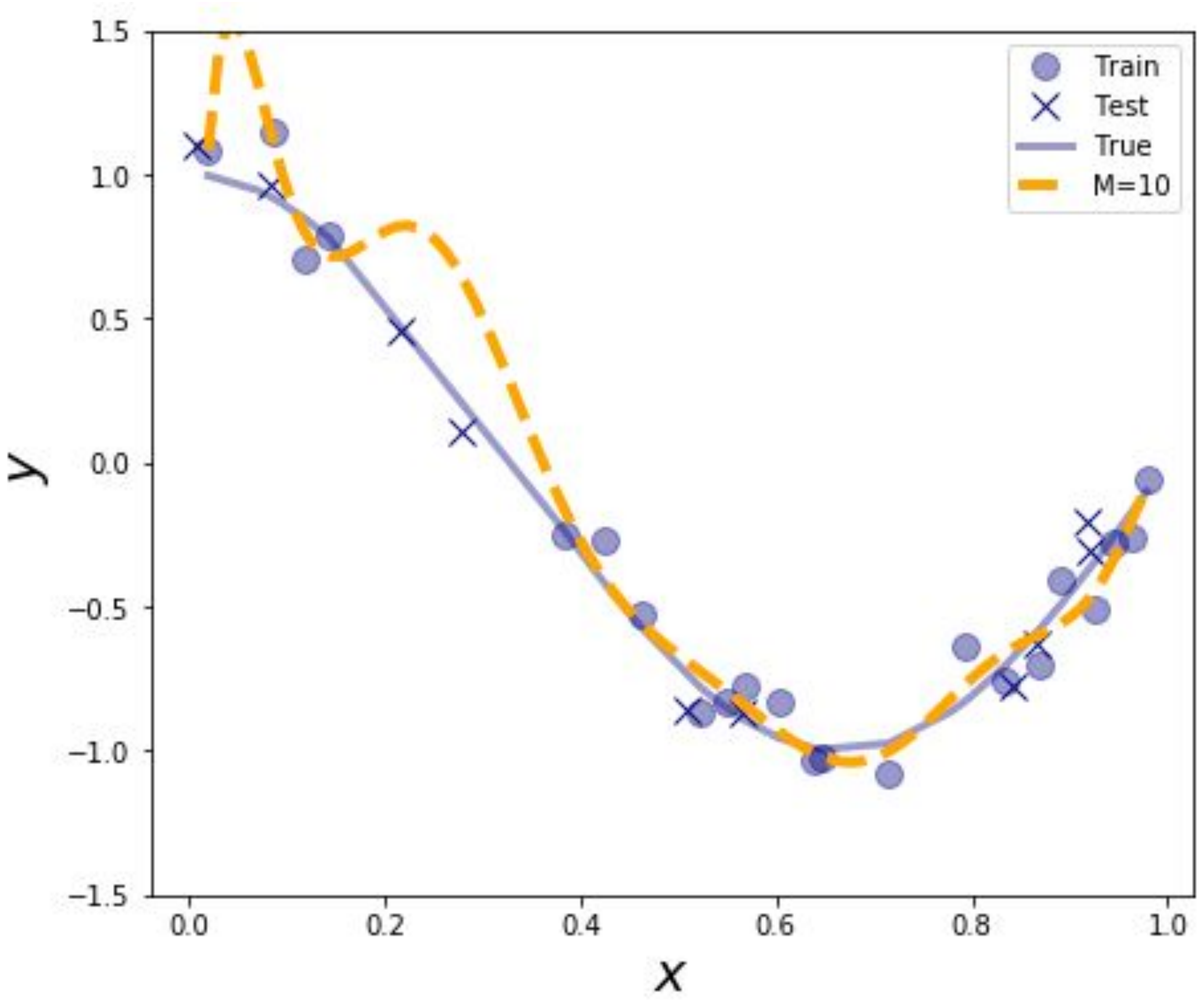
Aumenta el error de validación
comienza el overfitting…(M=8) ☹

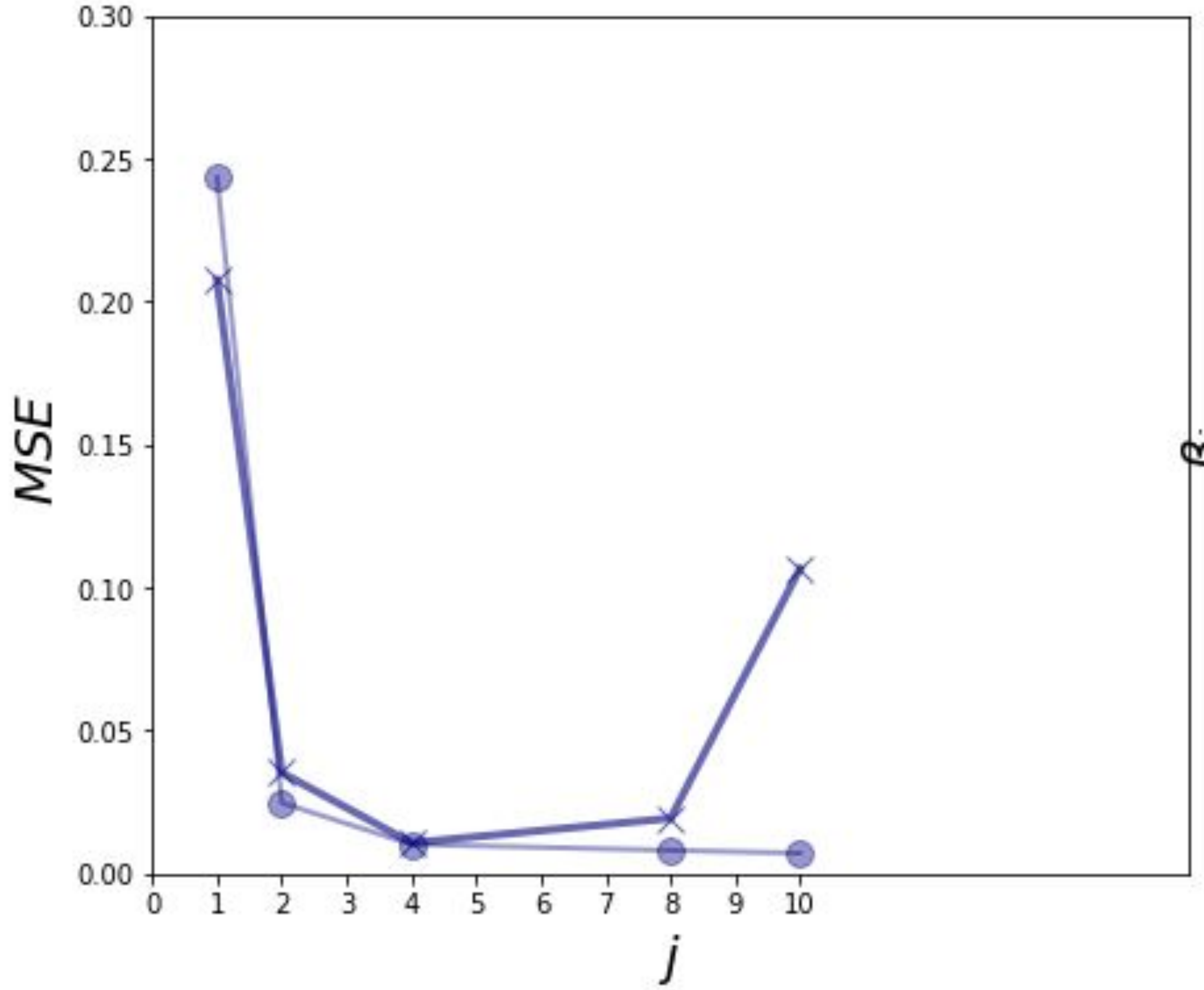
Valores de los coeficientes se hacen extremos ☹

REGRESIÓN Y OVERFITTING

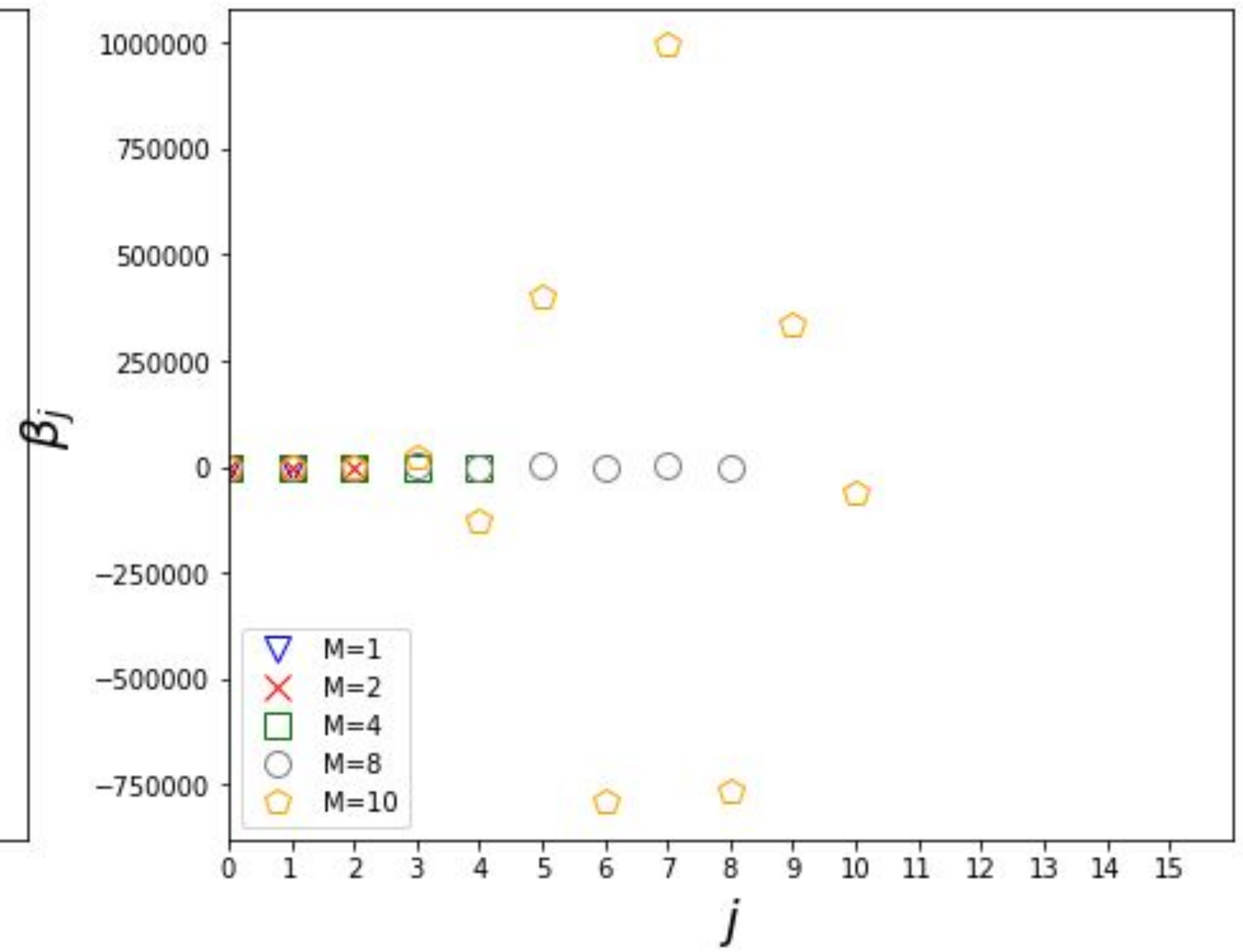
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M$$

Supongamos un modelo de regresión polinomial para un conjunto de **n=30** datos





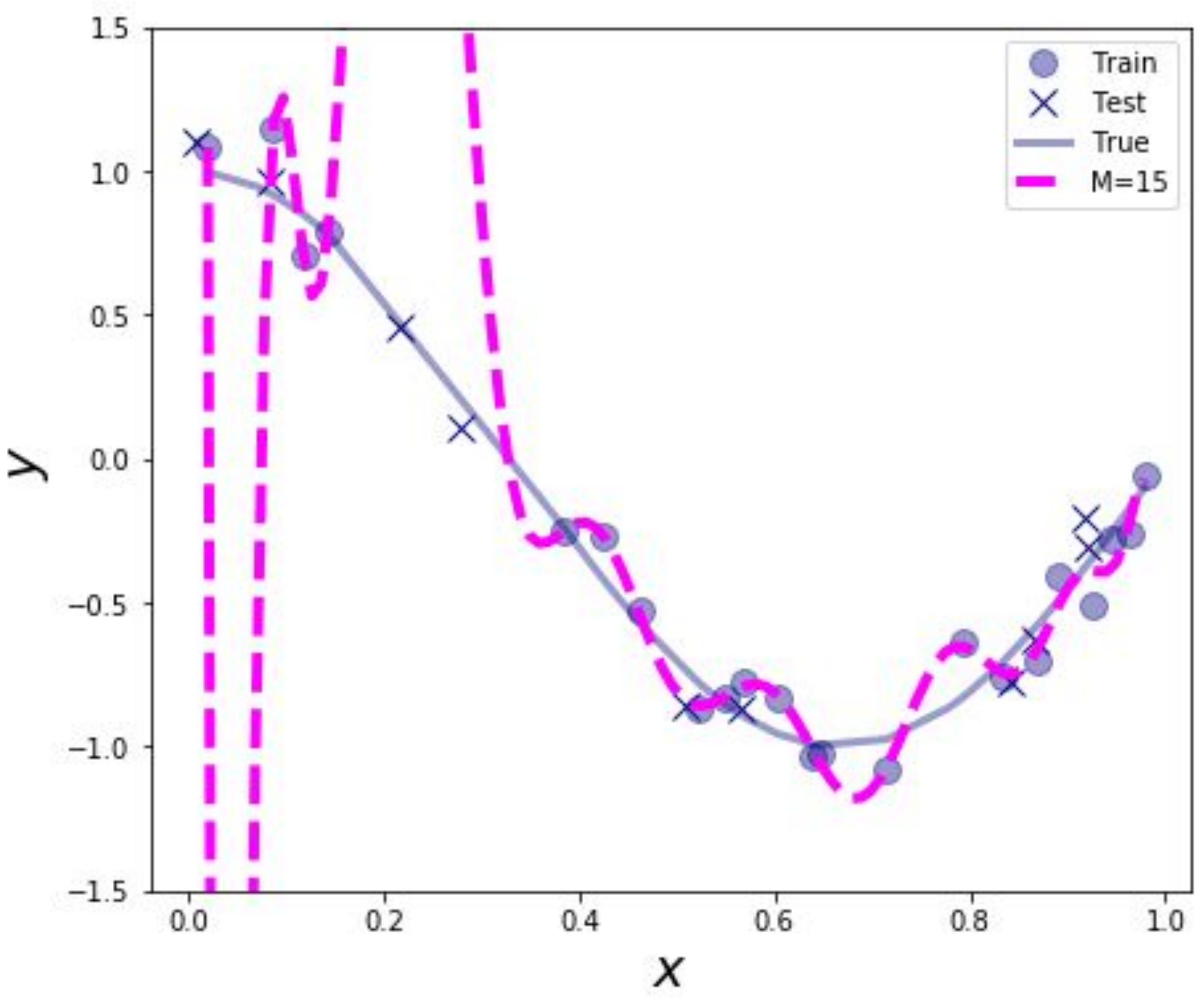
Overfitting!! ☹ ☹

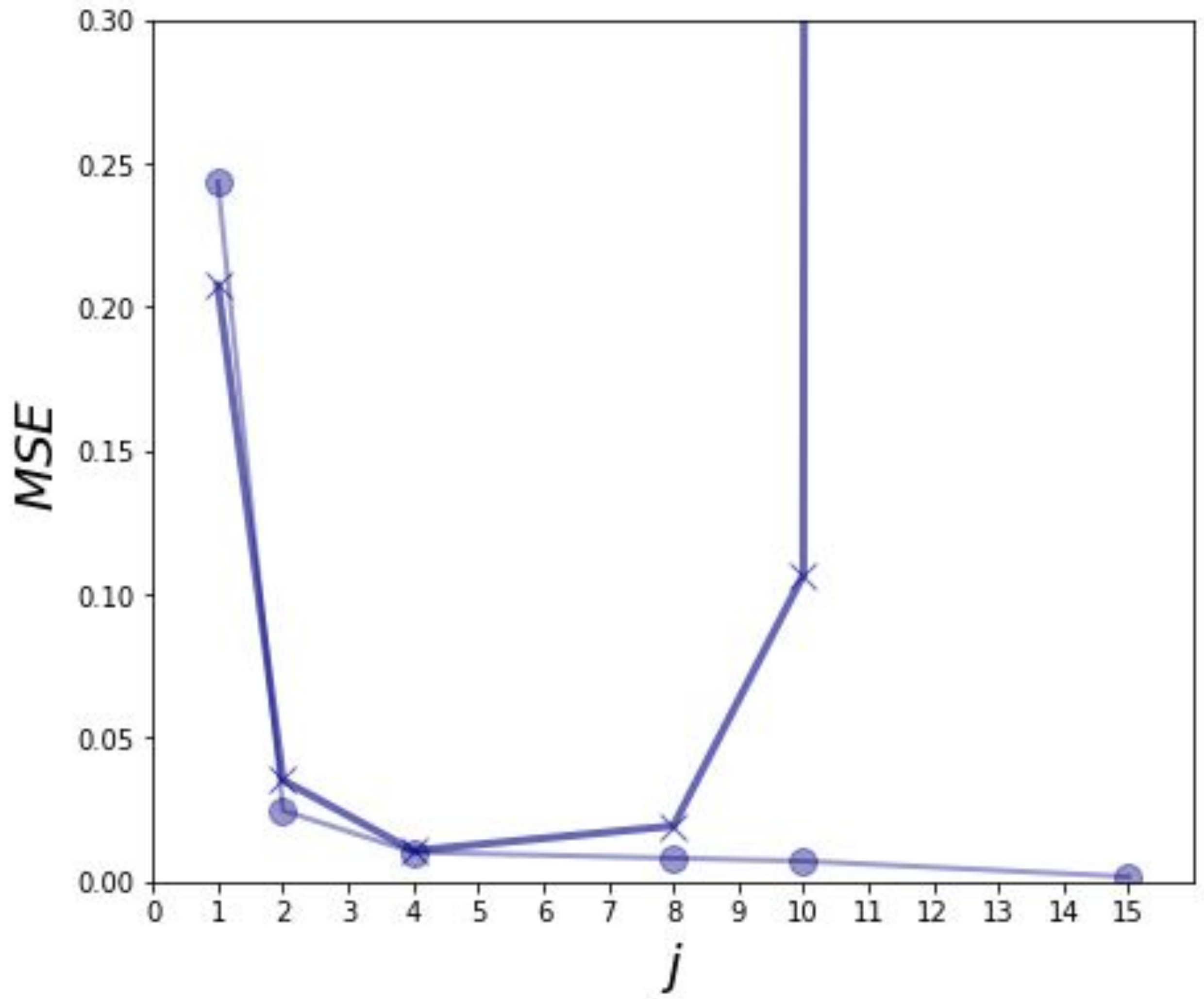


Valores de los coeficientes se hacen más extremos ☹

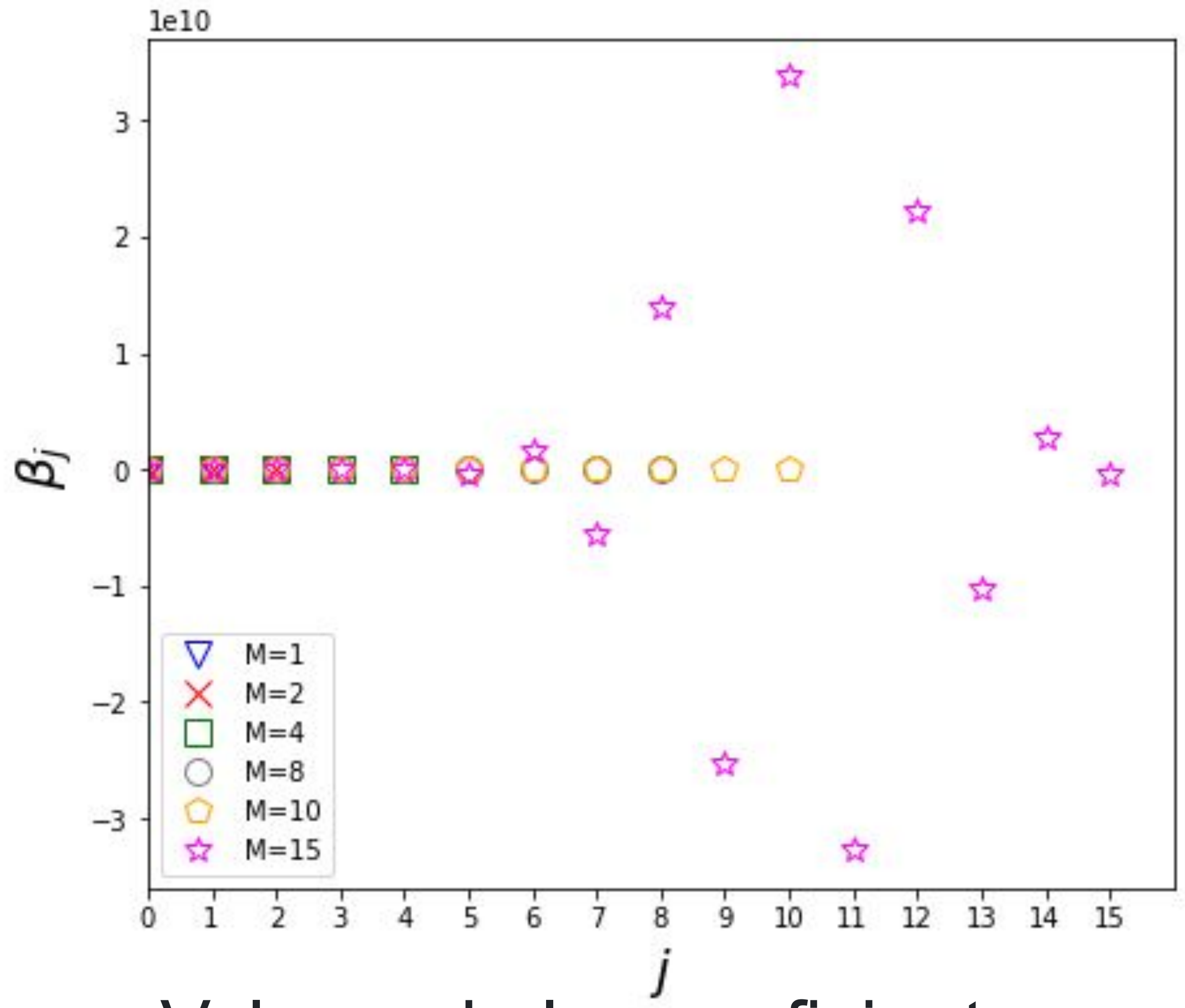
REGRESIÓN Y OVERFITTING

Supongamos un modelo de regresión polinomial para un conjunto de n=30 datos





Gran overfitting!! ☹️ ☹️ ☹️

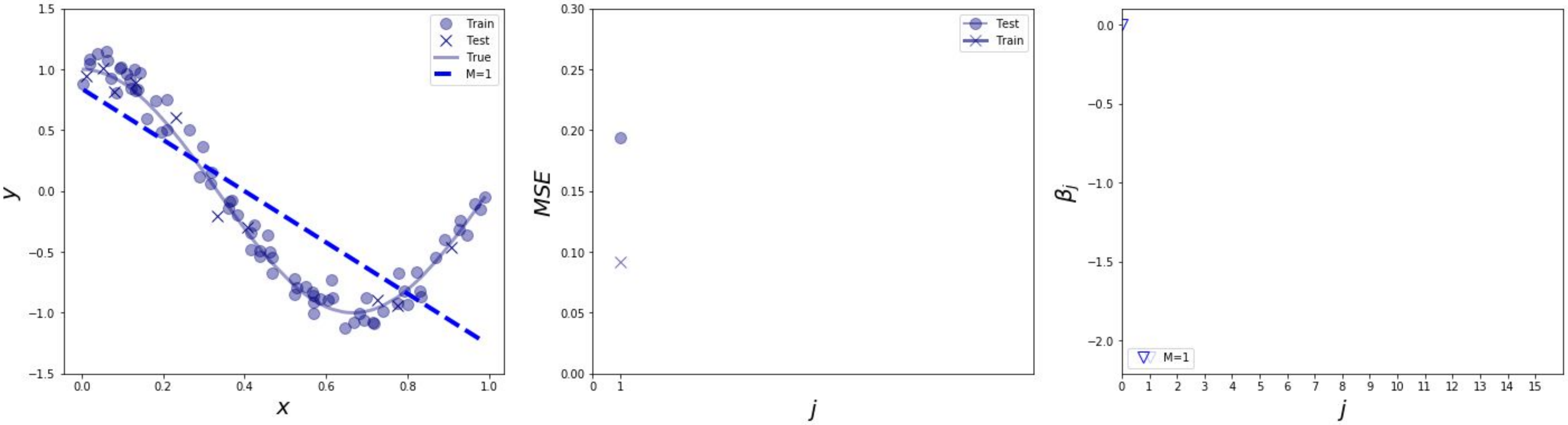


Valores de los coeficientes se hacen muy extremos ☹️

¿Qué pasa si agregamos más datos?

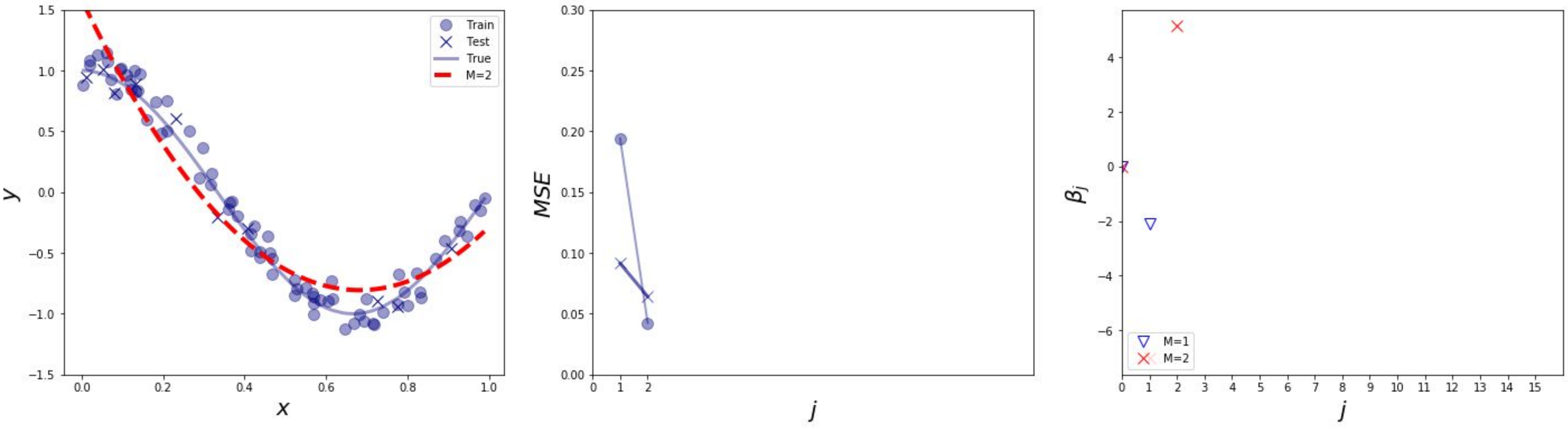
REGRESIÓN Y OVERFITTING

Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



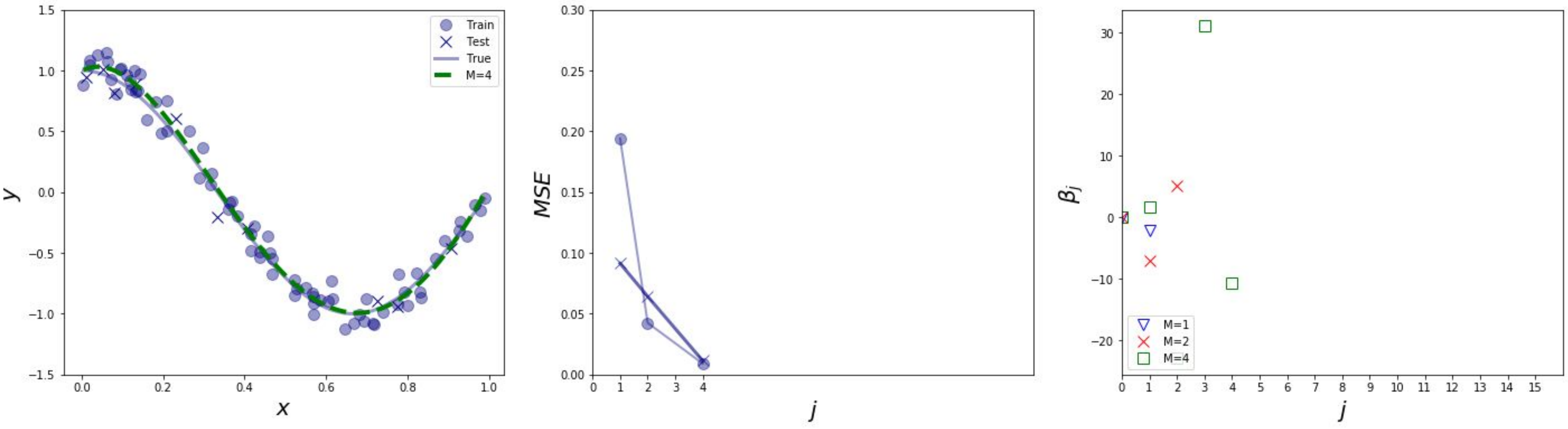
REGRESIÓN Y OVERFITTING

Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



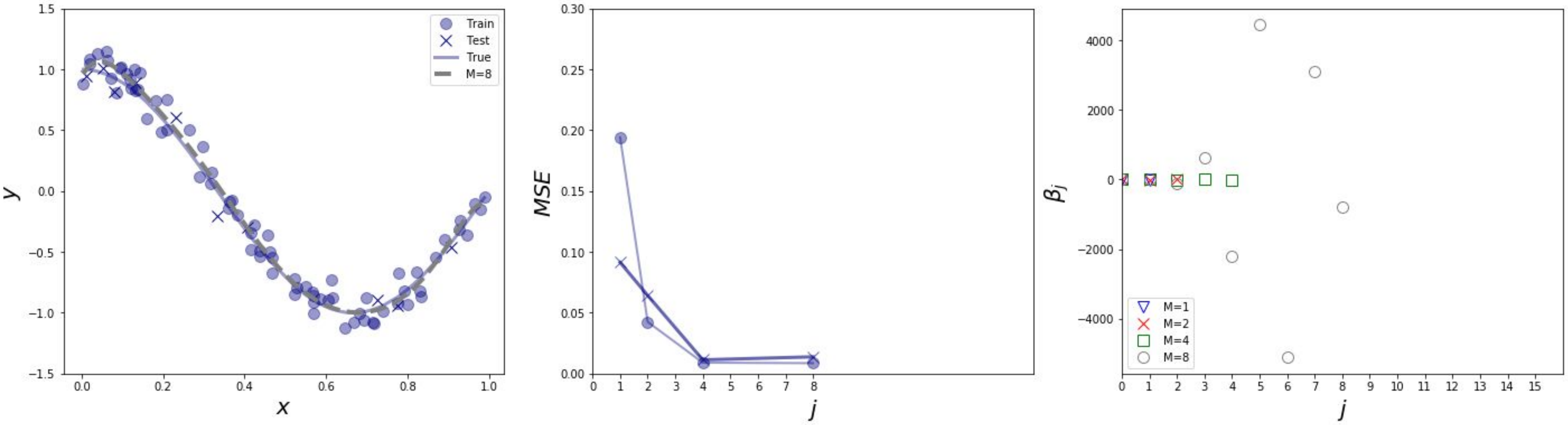
REGRESIÓN Y OVERFITTING

Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



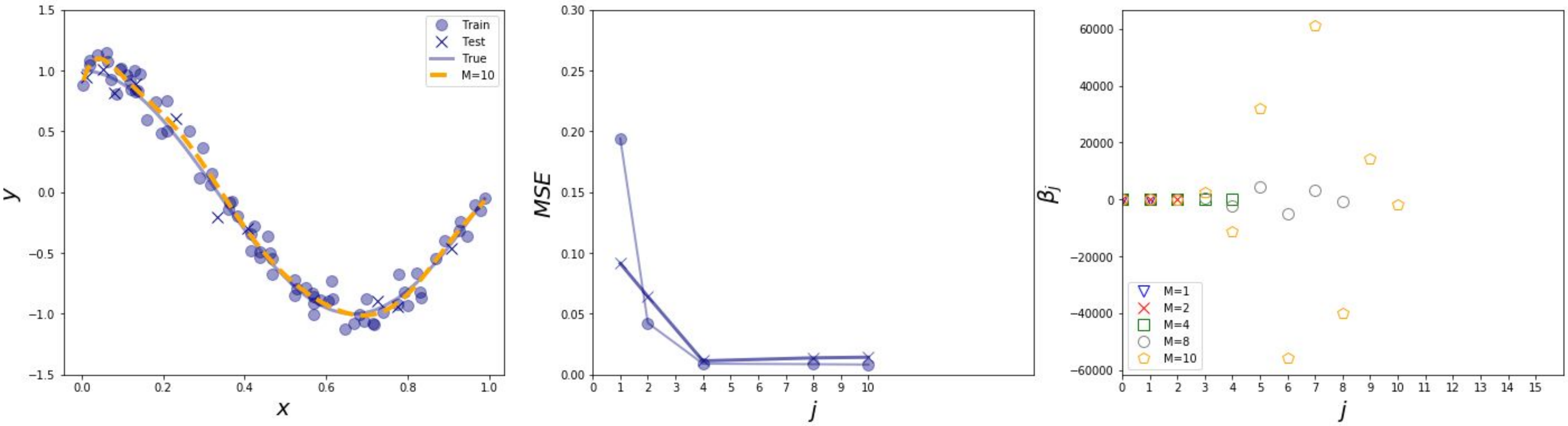
REGRESIÓN Y OVERFITTING

Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**

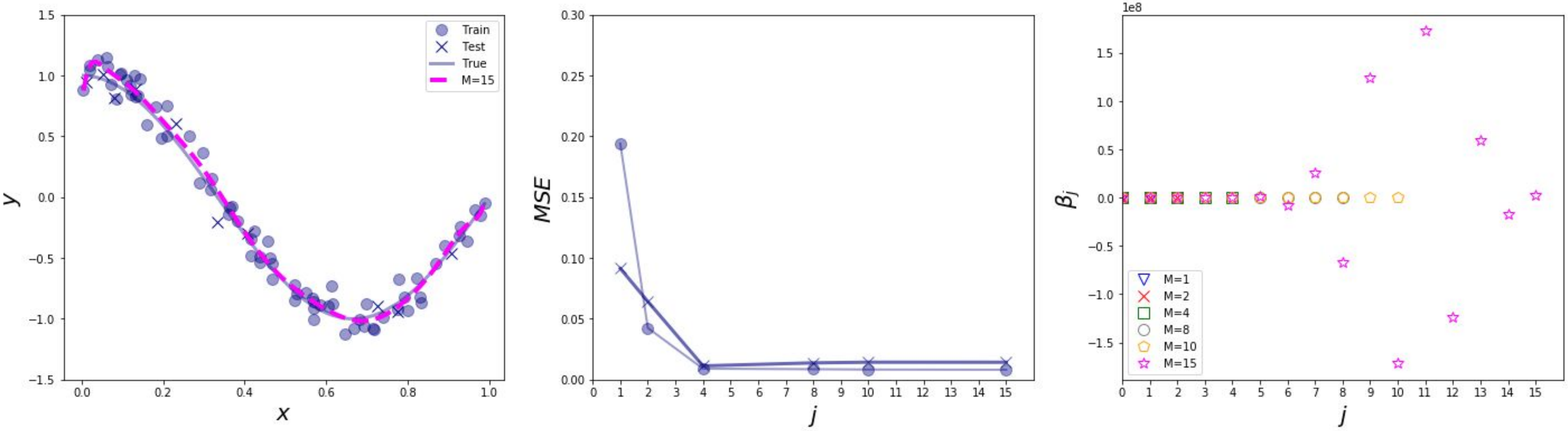


REGRESIÓN Y OVERFITTING

Supongamos un modelo de regresión polinomial para un conjunto de **n=100 datos**



REGRESIÓN Y OVERFITTING



□ Con más datos, puedo entrenar un modelo más complejo, con menor riesgo de caer en overfitting.

REGRESIÓN Y OVERFITTING

Conclusión: ¿Cómo evitar problemas de overfitting en modelos de regresión?

- Revisar la complejidad del modelo (grado de la función polinomial, cantidad de features)
- Agregar datos
- Reducir el ruido de los datos

- **Regularización:** agregar términos en la función de pérdida, que penalizan los valores extremos de los coeficientes de la regresión.

$$L_{reg}(\beta) = L(\beta) + \boxed{\alpha R(\beta)} \longrightarrow \text{regularización}$$

$$L_{LASSO}(\beta) = L(\beta) + \alpha \sum_{m=1}^{M} |\beta_m|$$

UC | Chile