

Curso de Inteligencia Artificial para Ingenieros

# Métodos de Regresión

Prof. Carlos A. Toro N.  
carlos.toro.ing@gmail.com  
2022

# Objetivos

Al finalizar esta presentación, usted podrá:

- ☐ Conocer los algoritmos más comunes de *Regresión*.
- ☐ Conocer algunas de las herramientas de programación típicas para este tipo de algoritmos.
- ☐ Conocer las métricas de evaluación de desempeño de estos algoritmos.

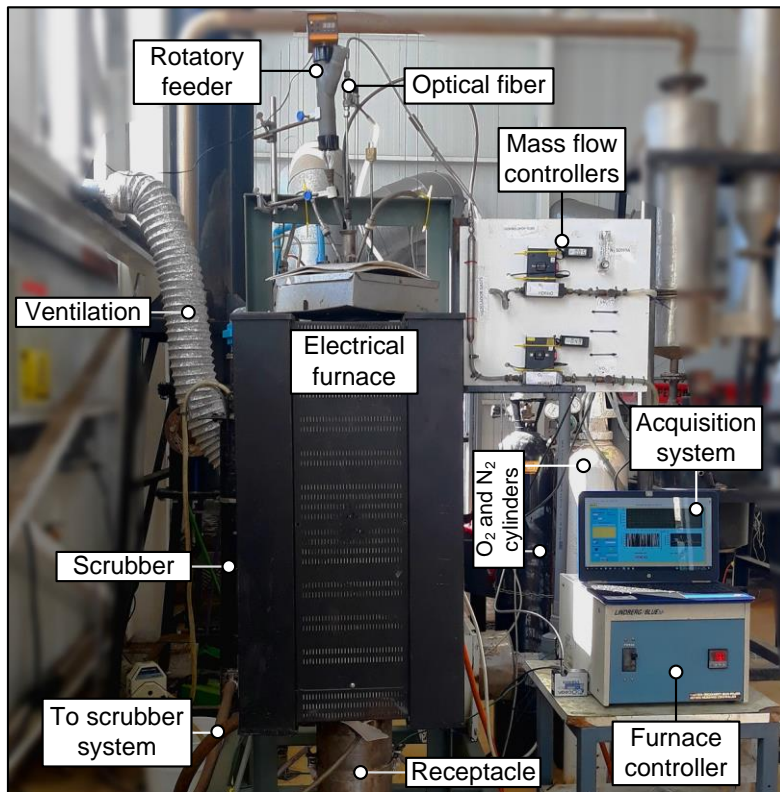
# Contenidos

- ☐ Previo: Datos
- ☐ Conceptos básicos de modelación
- ☐ Algoritmo de mínimos cuadrados
- ☐ Regresión Lineal
- ☐ Ajustes de curvas no lineales
- ☐ *Ejercicios prácticos*

**Previo: y los datos de donde salen?**

# Fuentes de datos

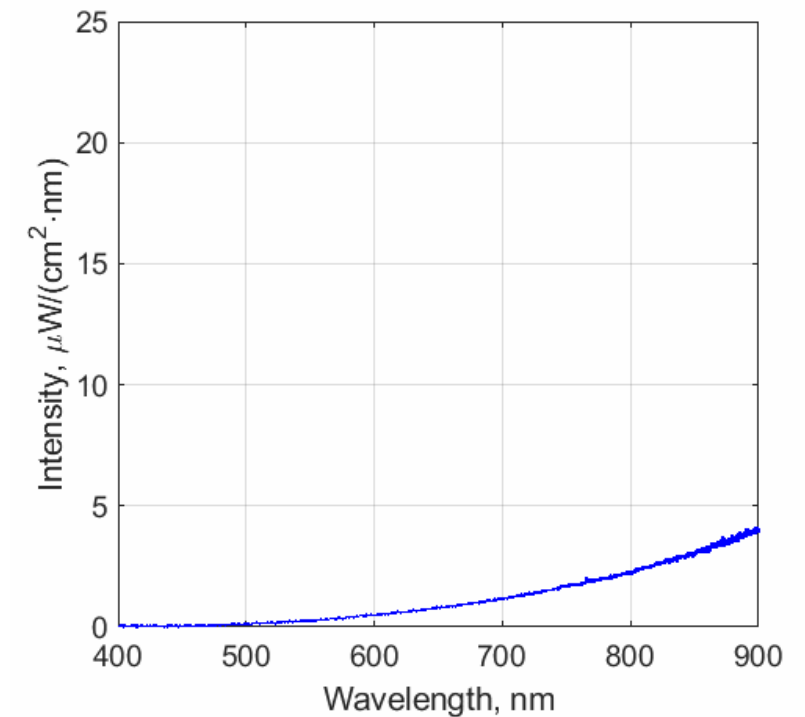
En nuestras especialidades, una de las principales fuentes de información son las **SEÑALES** capturadas desde procesos físicos o químicos por medio de sensores/instrumentos especiales, e.g.: Las siguientes imágenes muestran señales de radiación espectral capturadas desde un proceso de combustión de minerales de cobre y hierro en un montaje de laboratorio.



Montaje para experimentos de combustión



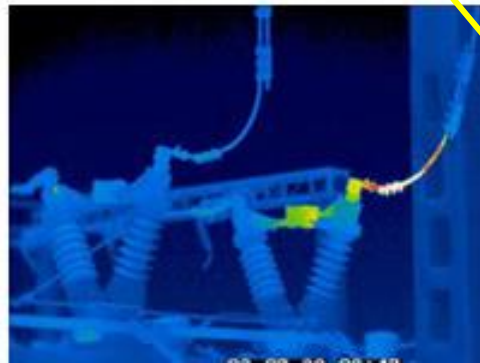
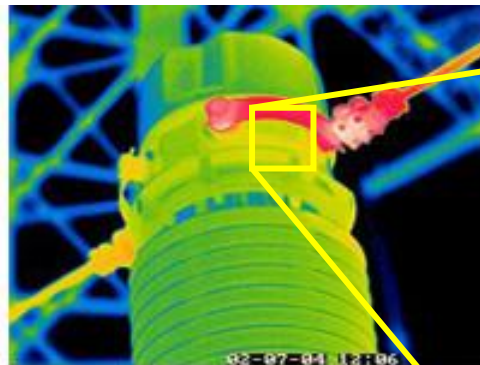
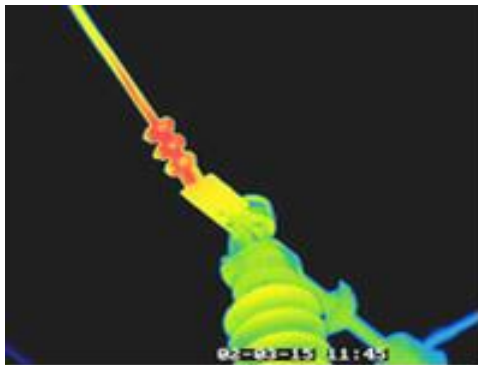
Llama generada



Señales capturadas

# Fuentes de datos

**Imágenes:** Un caso particular son las imágenes provenientes de cámaras, estas se interpretan en Matlab como arreglos de datos de: 1 matriz (caso monocromático), 3 matrices (caso color, una matriz para cada canal de color RGB), o más matrices (caso hiperespectral). Cada pixel de la imagen representa un valor de intensidad en una posición específica.

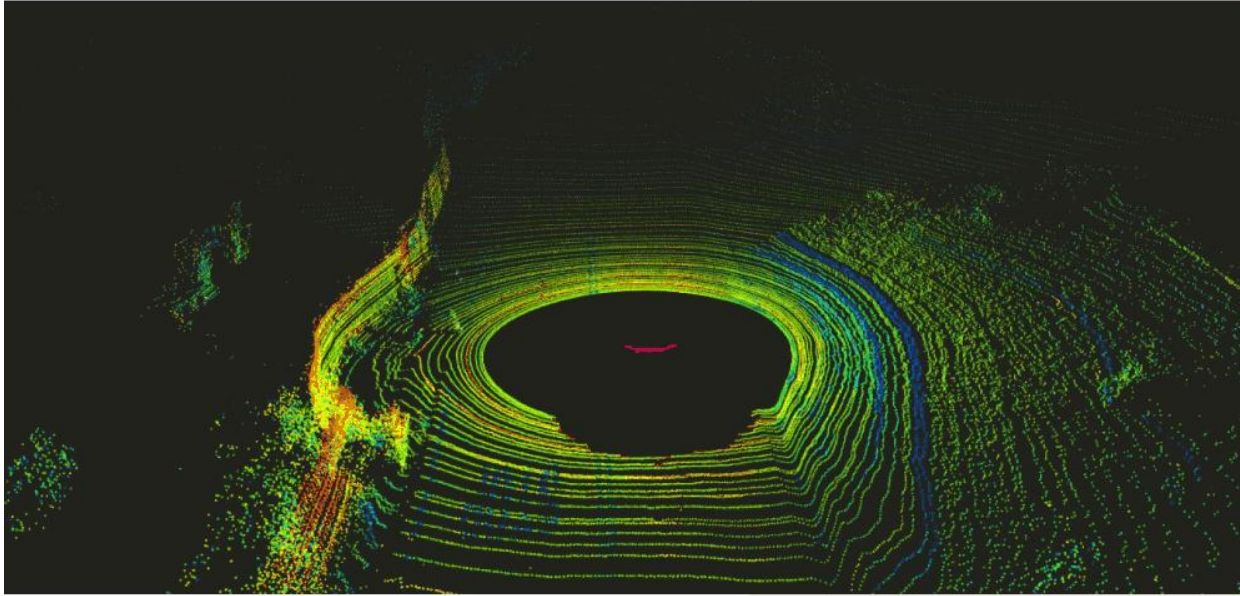


106	107	102	97	92	95
98	100	97	91	95	102
87	86	83	85	92	106
78	71	70	71	85	100
84	73	69	65	79	92
87	80	75	75	78	89
80	80	76	79	82	90
77	76	76	79	84	97
73	77	77	84	89	105
70	80	87	94	102	114
68	82	91	97	104	116

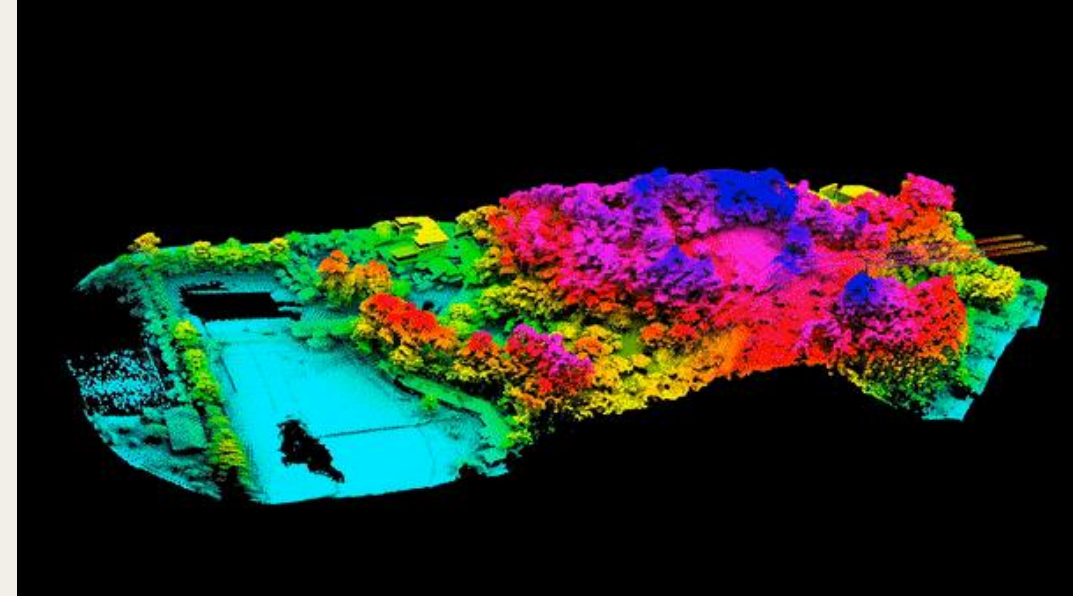
Termografías IR  
(imágenes en pseudocolor)

# Fuentes de datos

**Otras:** señales lidar (nubes de puntos 3D)

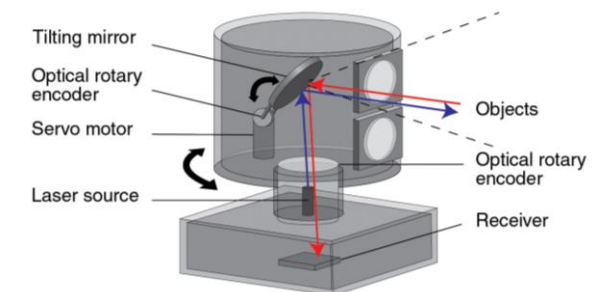


Mapeo especial con Lidar ([fuente](#))



Fotogrametría terrenos ([fuente](#))

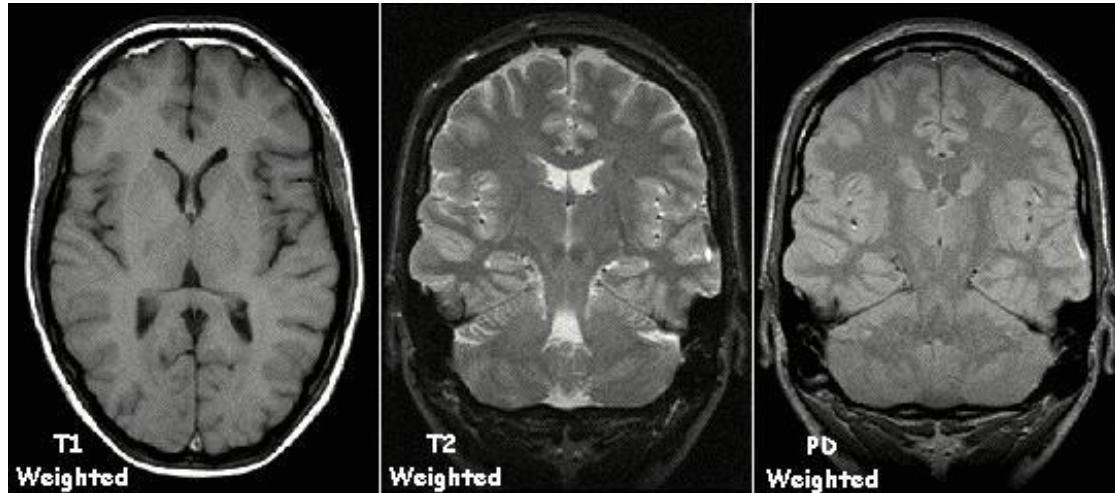
El principio de funcionamiento del LIDAR se basa en el método de tiempo de vuelo (Time Flight method). Mide la distancia en base a la diferencia de tiempo. Mide desde que sale disparado el laser, hasta que es detectado al regreso.



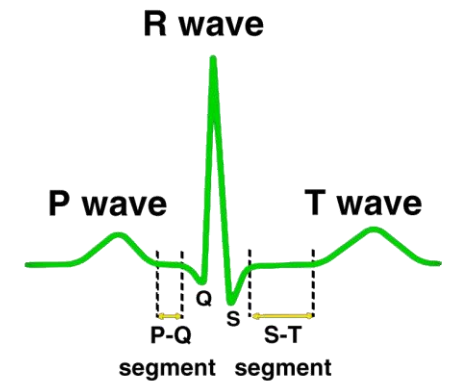
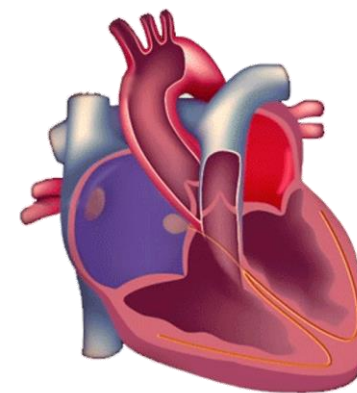
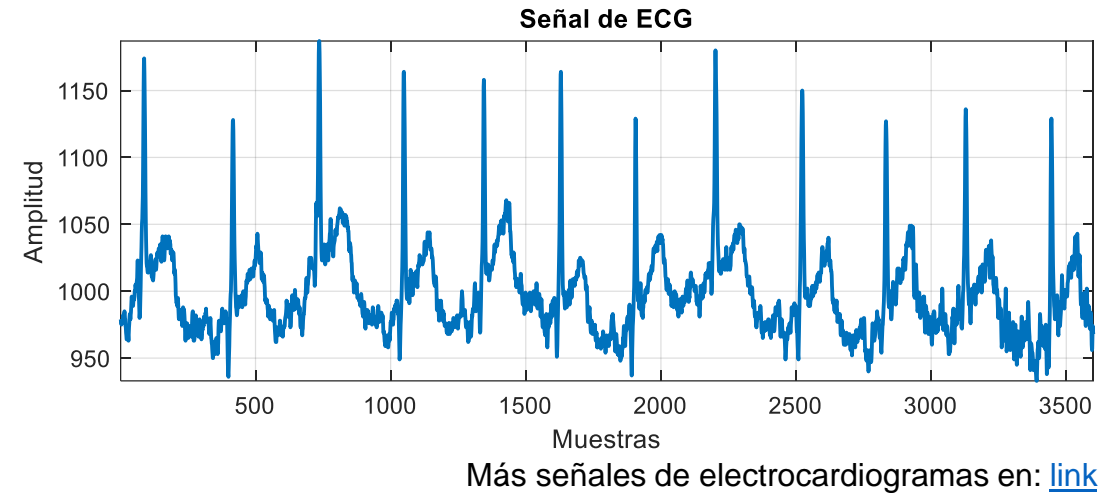
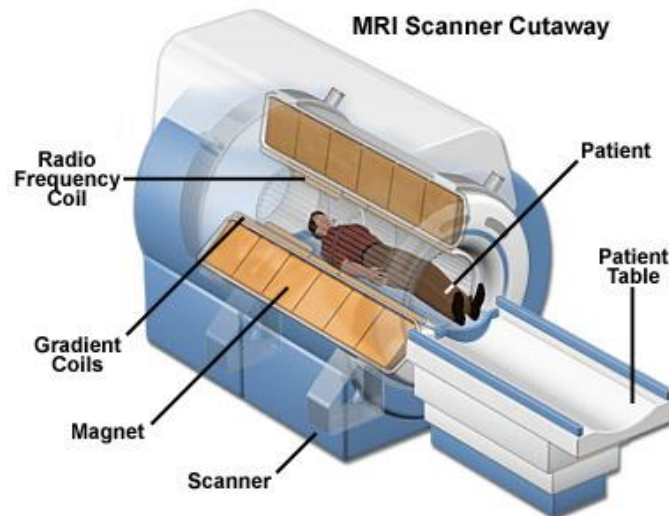


# Fuentes de datos

**Otras:** ECG (electrocardiogramas), EEG (electroencefalogramas), MRI (resonancia magnética), imágenes de rayos X, etc.



Resonancias magnéticas

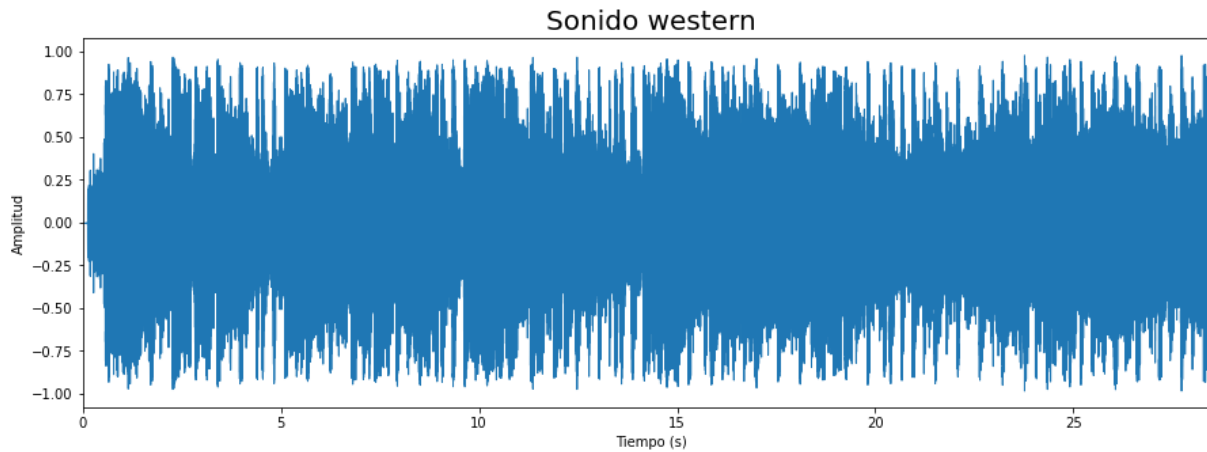


[Fuente del gif](#)



# Fuentes de datos

**Otras:** señales de audio

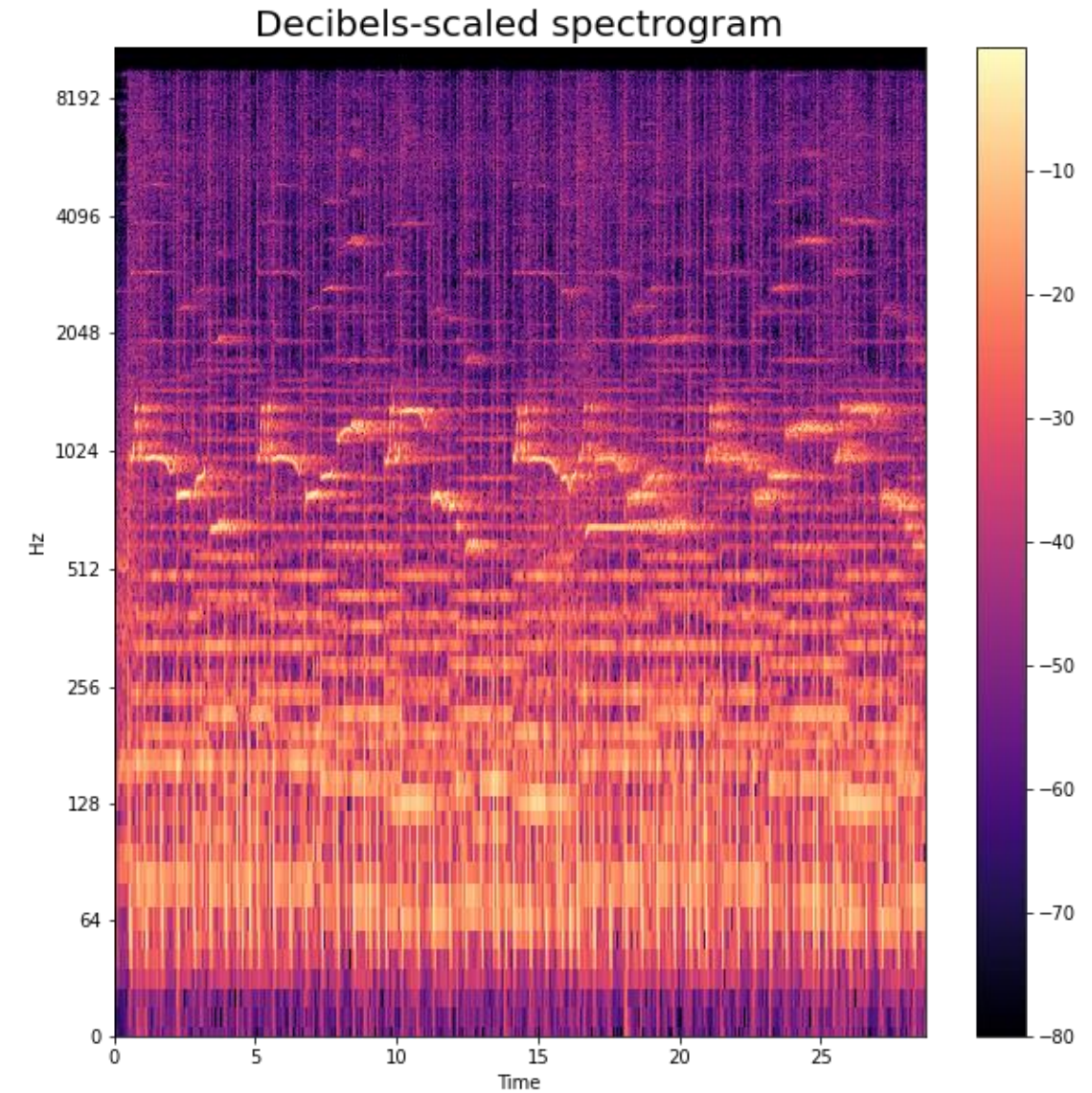


[Fuente audio](#)



0.9.2

Paquete de Python para análisis de audio y música: [enlace](#)



# Fuentes de datos

## Fuentes de datos disponibles en la web para practicar:

- Colección Wikipedia: [link](#)
- Fuente de datos kaggle.com: [link](#)
- Machine Learning Repository (UCI): [link](#)
- Datasets para Deep Learning en Matlab: [link](#)
- Datasets para Machine Learning en Matlab: [link](#)
- Datasets para Redes Neuronales en Matlab: [link](#)
- Datasets en Tensorflow: [link](#)
- IEEE Dataport: [link](#)

## Recordar:

- Antes de enfrentar un proyecto de desarrollo que involucre *Machine/Deep Learning* preguntar siempre por la **disponibilidad de los datos** y evaluar qué tan fácil sería obtenerlos si es que no existen aún.
- En **problemas multidisciplinarios** con otras áreas (ej. geología, biología, química, astronomía, mecánica, ... ), asumir que no sabemos todo acerca del **análisis particular de las señales involucradas**, se puede llegar a conclusiones erróneas si no se colabora con los expertos del área (ustedes) o se estudia correctamente el problema.

# Regresión: Conceptos Básicos

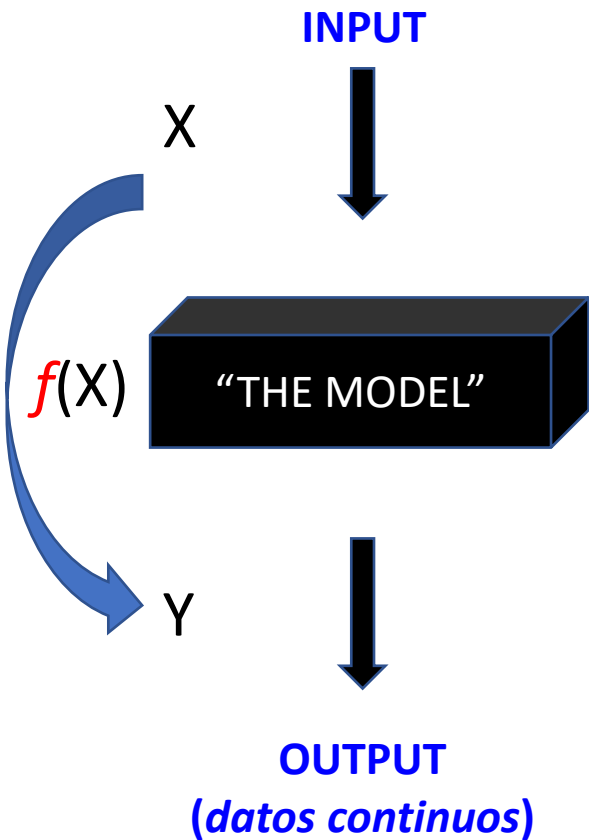
# Motivación

**Tres tipos de modelos:** *Experimentales*, *teóricos* (ej. basados en leyes físicas) y *modelos mixtos*. (Más adelante se verán otras clasificaciones).

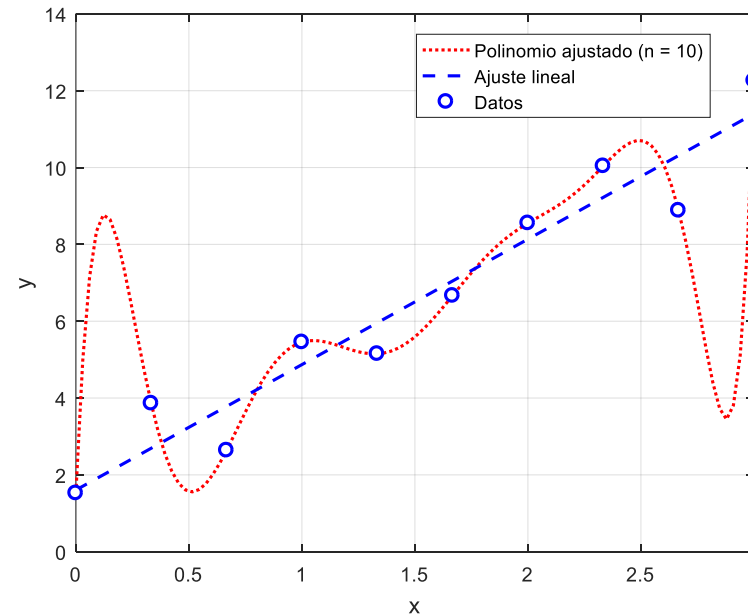
## ❑ Para qué modelar?:

- Diseño;
- Control de procesos (ej. Identificación de parámetros de sistemas dinámicos);
- Optimización;
- Evaluación/planificación de experimentos;
- Búsqueda y solución a fallas y diagnósticos;
- Determinar cantidades que no pueden ser medidas directamente;
- Simular en lugar de llevar a cabo experimentos costosos;
- Estudios de factibilidad para determinar el potencial de una solución antes de construir un equipo prototipo o dispositivo;
- Analizar condiciones de proceso que serían riesgosas de implementar en la práctica, etc.

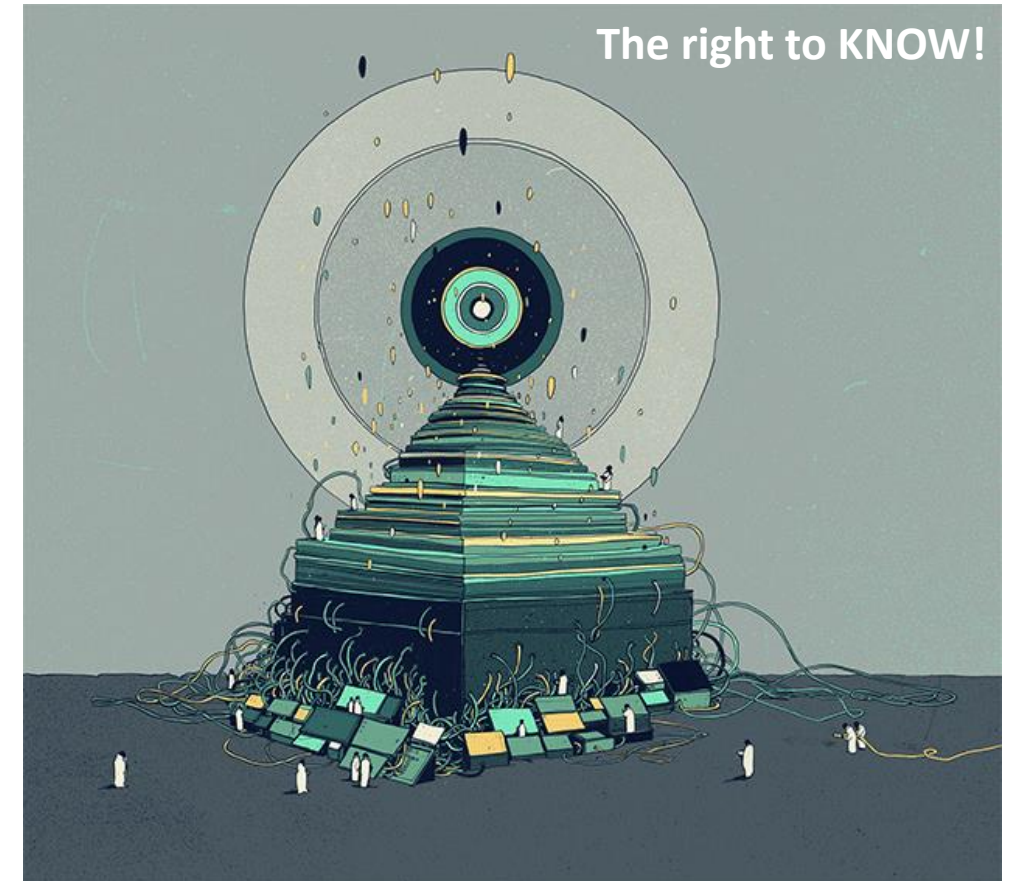
# Motivación



Enfoque de Caja Negra. Principalmente basado en modelación a partir de **DATOS**



Ejemplo: Ajuste polinomial



Del artículo: *Can we open the black box of AI?*

[source, DOI: 10.1038/538020a]

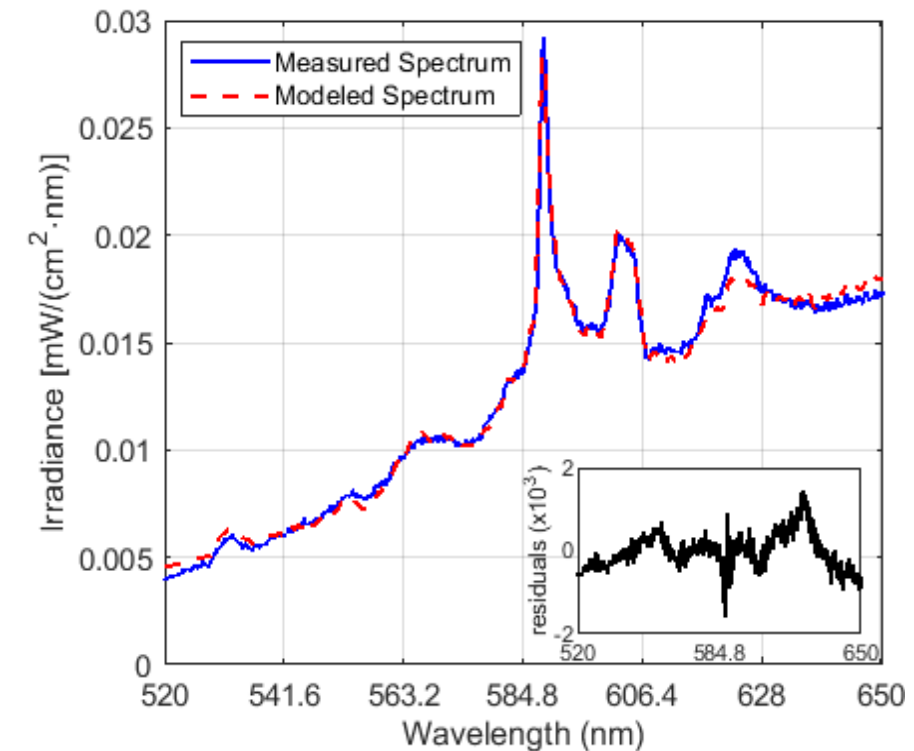
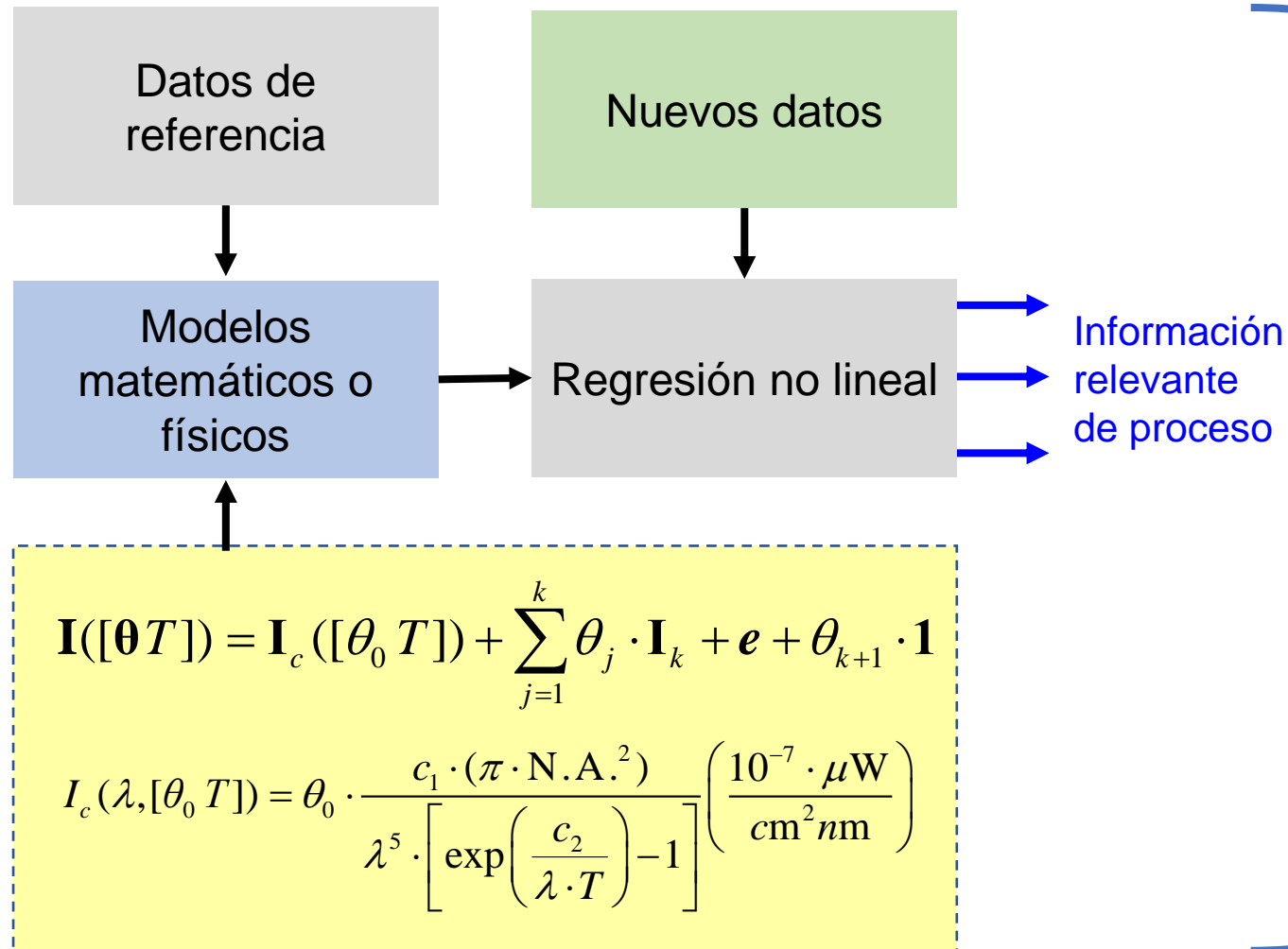
Principalmente encontrado en aplicaciones de ciencias experimentales.

**Ojo:** Aún cuando obtengamos buenos resultados al modelar, muchas aplicaciones requieren explicaciones y transparencia, e.g. medicina.



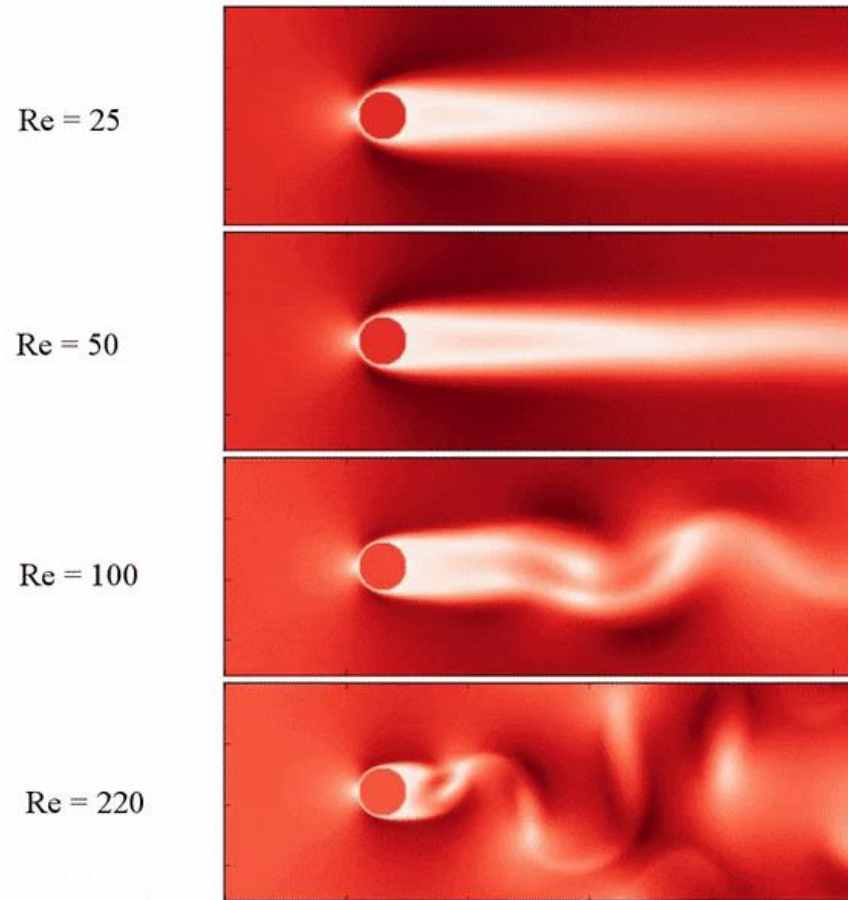
# Motivación

Ejemplo de un enfoque de modelación mixta ([ref. del ejemplo](#)):



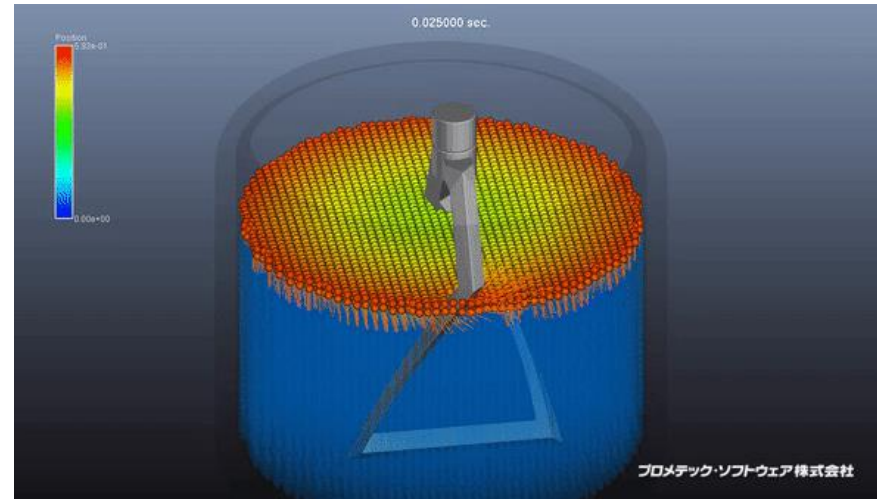
Modelación analítica + empírica  
(modelos de caja gris)

# Motivación



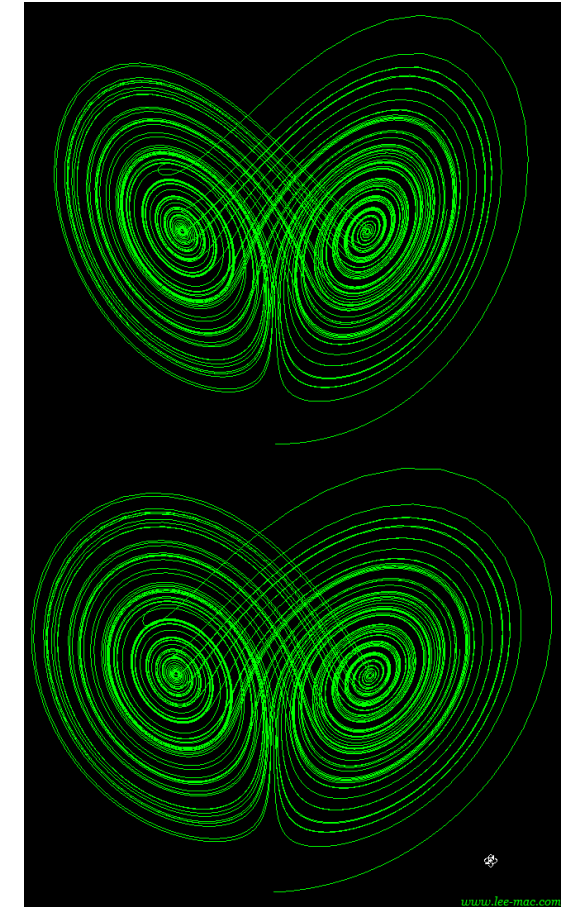
Efecto del número de Reynolds en el flujo alrededor de un cilindro.

[source: <https://www.youtube.com/watch?v=8WtEuW0GLg0>]



CFD particles simulation [source: Solidworks simulation  
<https://enginsoftusa.com/SOLIDWORKS-CFD-Software.html> ]

Modelos fenomenológicos, modelos de caja blanca. Principalmente basados en leyes químicas y/o físicas.

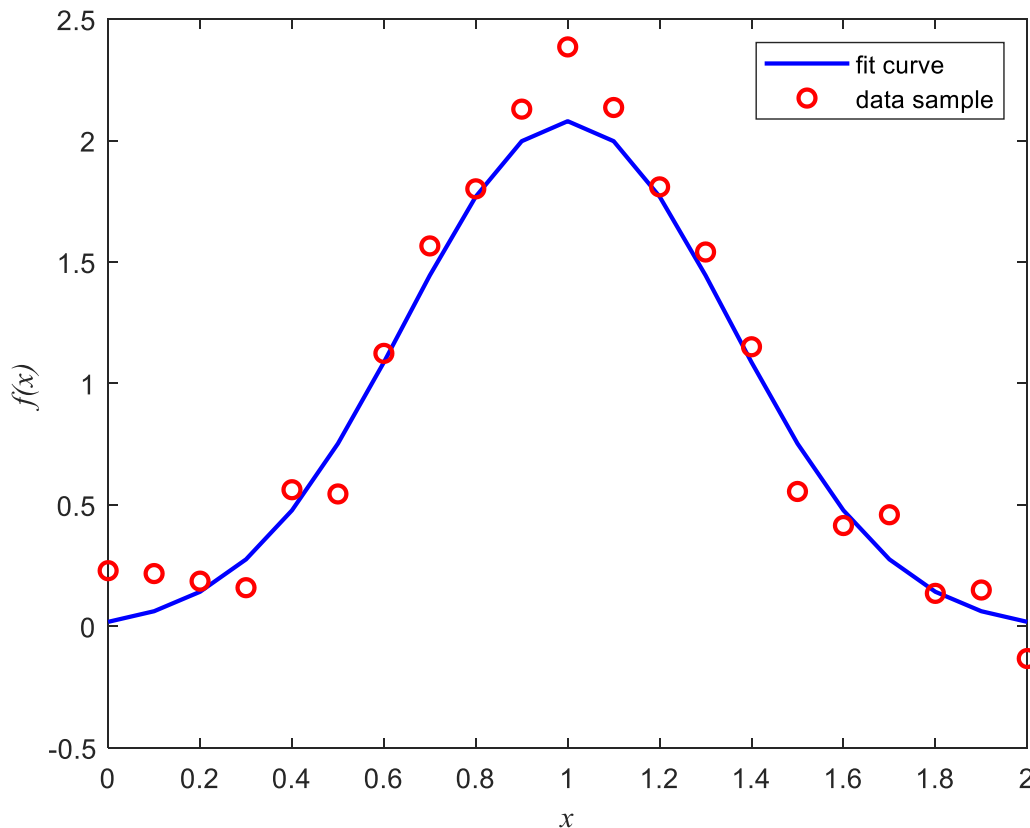


Lorenz attractor, basis for weather analysis and CFD simulations

[source: <http://www.lee-mac.com/attractors.html>]

# Regresión

Los problemas de regresión pueden ser vistos como problemas de ajustes de funciones, consiste simplemente en encontrar un set de parámetros/coeficientes para que una ecuación se ajuste a un conjunto de datos (**continuos**) medidos.



$$f(x) = A + Be^{-\left(\frac{x-C}{D}\right)^2}$$

“El ajuste de curvas” determina los valores de  $A$ ,  $B$ ,  $C$  y  $D$  tal que  $f(x)$  representa de la mejor forma a los datos, en este caso se obtuvo:

$$A = -0.02;$$

$$B = 2.1;$$

$$C = 1;$$

$$D = 0.5;$$

# Regresión

Y qué modelo usamos?

❑ **2 opciones. O sabemos el modelo a priori o no!**

❑ En caso de no saber el tipo de modelo o función que mejor se ajusta a una tabla de datos del tipo  $(x_i, y_i)$ ,  $i = 1, \dots, M$  (no necesariamente solo una variable dependiente e independiente), **se recomienda la visualización de los datos para intuir algún tipo de función**. Algunas funciones/modelos clásicos son los siguientes:

▪ Modelos lineales

$$Ej : y = ax + b$$

▪ Polinomiales

$$Ej : y = ax^6 + bx^3 + cx + d$$

▪ Exponenciales

$$Ej : y = a + Ae^{Bx}, \quad y = a(1 - e^{-bx})$$

▪ Sigmoidales

$$Ej : y = a / \left[ 1 + be^{-kx} \right]$$

▪ De potencia

$$Ej : y = a \cdot (x - b)^c$$

▪ Trigonométricos,

$$Ej : y = A \cdot \text{sen}(2 \cdot \pi Bt + C)$$

▪ Etc.

# Regresión: Caso lineal general (caso 1 var in – 1 var out)

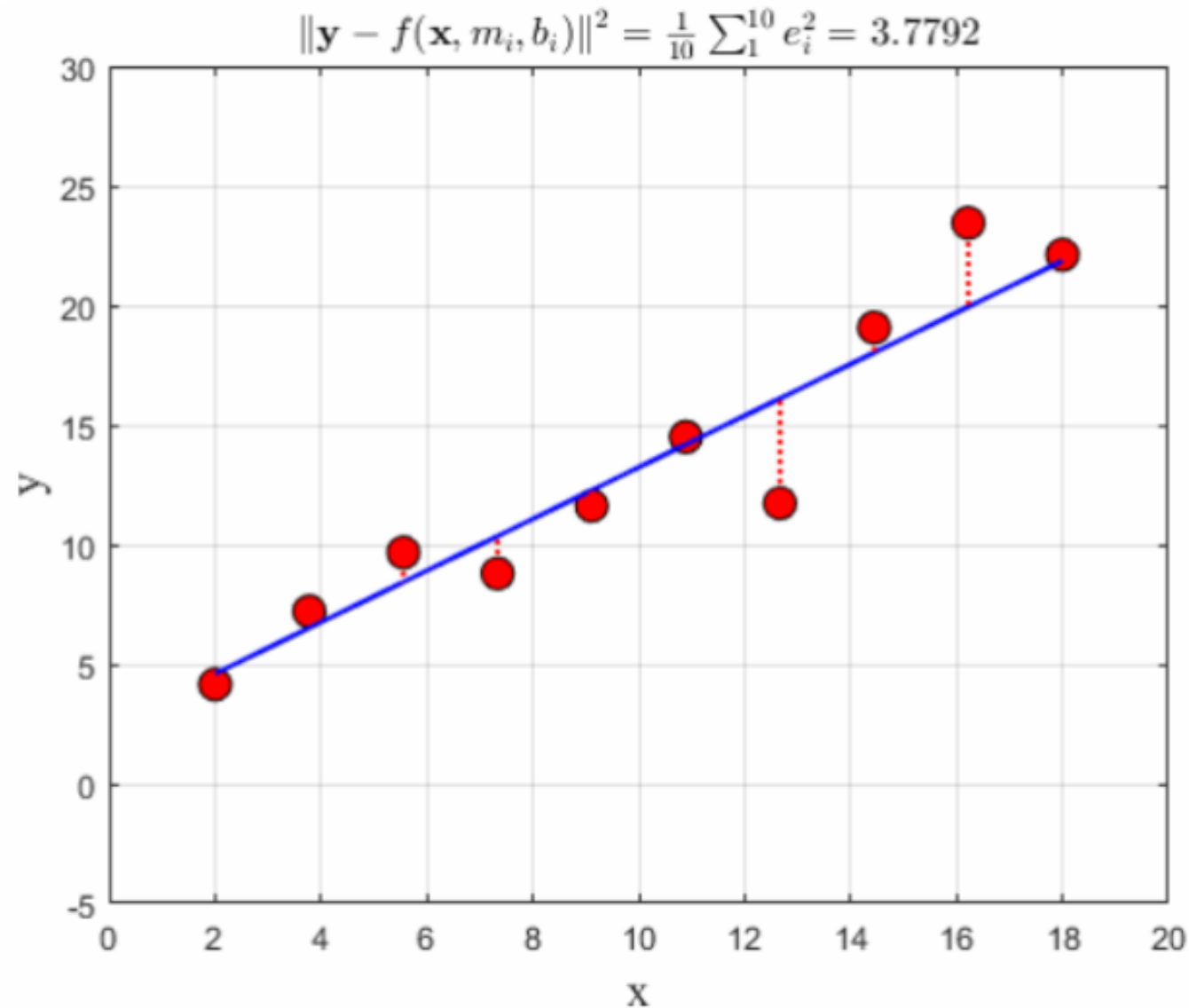
## Ajustar el modelo

$$f(x_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 g_1(x_i) + \dots + \beta_{N-1} g_{N-1}(x_i) + \varepsilon$$

a un conjunto de  $M$  mediciones experimentales (datos) del tipo :  $(x_1, y_1), \dots, (x_M, y_M)$  , donde:

- $\boldsymbol{\beta} = \beta_0, \beta_1, \dots, \beta_{N-1}$  , son los  $N$  parámetros de regresión a encontrar,  $M > N$
- $x_i$  : variable independiente, predictor, entrada, regresor, característica, variable observada, variable explicadora.
- $y_i$  : variable dependiente, resultado, objetivo, etiqueta, salida, variable de respuesta.
- $g_i$  : funciones base combinadas linealmente con los parámetros.
- $\varepsilon$  : *el error se asume  $N(\mu = 0, \sigma^2)$* , verificar esta suposición al final del proceso de estimación, si la condición no se satisface, algoritmos más robusto o transformaciones sobre los datos podrían ser requeridos.

# Regresión: Caso lineal general (caso 1 var in – 1 var out)



Cómo encontramos la mejor línea recta?

Una forma:

**Mínimos Cuadrados !**



# Regresión: Mínimos Cuadrados Ordinarios

En forma matricial podemos escribir el problema de estimación como:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} 1 & g_1(x_1) & \cdots & g_{N-1}(x_1) \\ 1 & g_1(x_2) & \cdots & g_{N-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & g_1(x_M) & \cdots & g_{N-1}(x_M) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{N-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix}$$

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad M > N$$

Notar que este es un sistema lineal sobredeterminado ( $M > N$ ), donde la solución puede ser estimada como:

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$$

Esta solución se conoce como la solución en el sentido de los mínimos cuadrados ordinarios (OLS) del problema de regresión, ya que minimiza los residuos  $\|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|^2$

# Regresión: Métricas de Calidad del Ajuste

## ❑ Métricas de calidad de ajuste

- Útiles para saber si el ajuste propuesto a los datos es adecuado
- Útiles al comparar y seleccionar modelos
- Entre estas encontramos:
  - Valor RMSE (*Root Mean Square Error*)
  - Coeficiente de determinación  $R^2$
  - SMAPE (*Symmetric Mean Absolute Percentage Error*)
- A continuación se describen estas métricas teniendo en cuenta un conjunto de  $N$  muestras de una función  $y_i = f(x_i)$ ,  $i = 1, \dots, N$  y un ajuste de esa función descrito por

# Regresión: Métricas de Calidad del Ajuste

## ❑ Valor **RMSE** (*Root Mean Square Error*, Raíz Cuadrada del Error Cuadrático Promedio)

- Mientras mas cercano sea su valor a cero, mejor será el ajuste de la función o el modelo a los datos.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left[ y_i - \hat{f}(x_i) \right]^2}$$

## ❑ **SMAPE** (*Symmetric Mean Absolute Percentage Error*, Error Porcentual Absoluto Promedio y Simétrico)

$$SMAPE = \frac{\sum_{i=1}^N |y_i - \hat{f}(x_i)|}{\sum_{i=1}^N y_i + \hat{f}(x_i)} \cdot 100\%$$

- Su valor será del 0% para un ajuste perfecto, y cercano al 100% para el peor de los casos.
- Es de más fácil interpretación que el valor RMSE.
- Nos proporciona una cota superior e inferior de valores.

# Regresión: Métricas de Calidad del Ajuste

## ❑ *Coeficiente de determinación $R^2$*

- El coeficiente de determinación  $R^2$  (o coeficiente de correlación  $R$ ), es una medida de que tan bueno fue el ajuste de los datos al modelo propuesto.
- En estadística, este coeficiente indica la proporción de la varianza en la variable dependiente que es predecible por la(s) variable(s) independiente(s).
- Toma valores entre 0 y 1. Donde  $R^2 = 1$  indica una correlación perfecta, y  $R^2 = 0$  indica que no existe correlación entre las variables. En este caso, entre los datos  $y_i$  originales y los valores predichos con el modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^N [y_i - \hat{f}(x_i)]^2}{\sum_{i=1}^N [y_i - \bar{y}]^2} \quad \left. \vphantom{\sum_{i=1}^N} \right\} \text{ Con: } \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$$

# Regresión

**Otra forma de ver el planteamiento del problema de regresión es mediante el concepto de optimización.**

- ❑ En un problema de optimización tratamos de encontrar los parámetros óptimos asociados a una función de costo o de pérdida (*loss function*), en el campo del *machine learning* podemos plantear un modelo de regresión lineal de la siguiente forma, donde  $w$  indica el parámetro de ponderación (pendiente para el caso univariado),  $b$  indica el sesgo (coeficiente de posición en el caso univariado) y  $x$  indica el vector de características de la muestra, predictor o variable independiente.

$$h_w(x) = w^T x + b$$

- ❑ La relación entre el valor predicho por el modelo y el valor real es la siguiente, donde  $y$  indica el valor real y  $\varepsilon$  indica el error.

$$y = w^T x + b + \varepsilon$$

- ❑ El error  $\varepsilon$  está influenciado por muchos factores de forma independiente. Según el teorema del límite central, el error  $\varepsilon$  sigue una distribución normal. Luego, **la función de pérdida** de la regresión lineal es la siguiente:

$$J(w) = \frac{1}{2m} \sum (h_w(x) - y)^2$$

- ❑ Para que el valor previsto se acerque al valor real, debemos minimizar el valor de pérdida. Podemos usar el método de descenso de gradiente (lo veremos en detalle en la unidad de redes neuronales) para calcular el parámetro  $w$  cuando la función de pérdida alcanza el mínimo y luego completar la construcción del modelo.

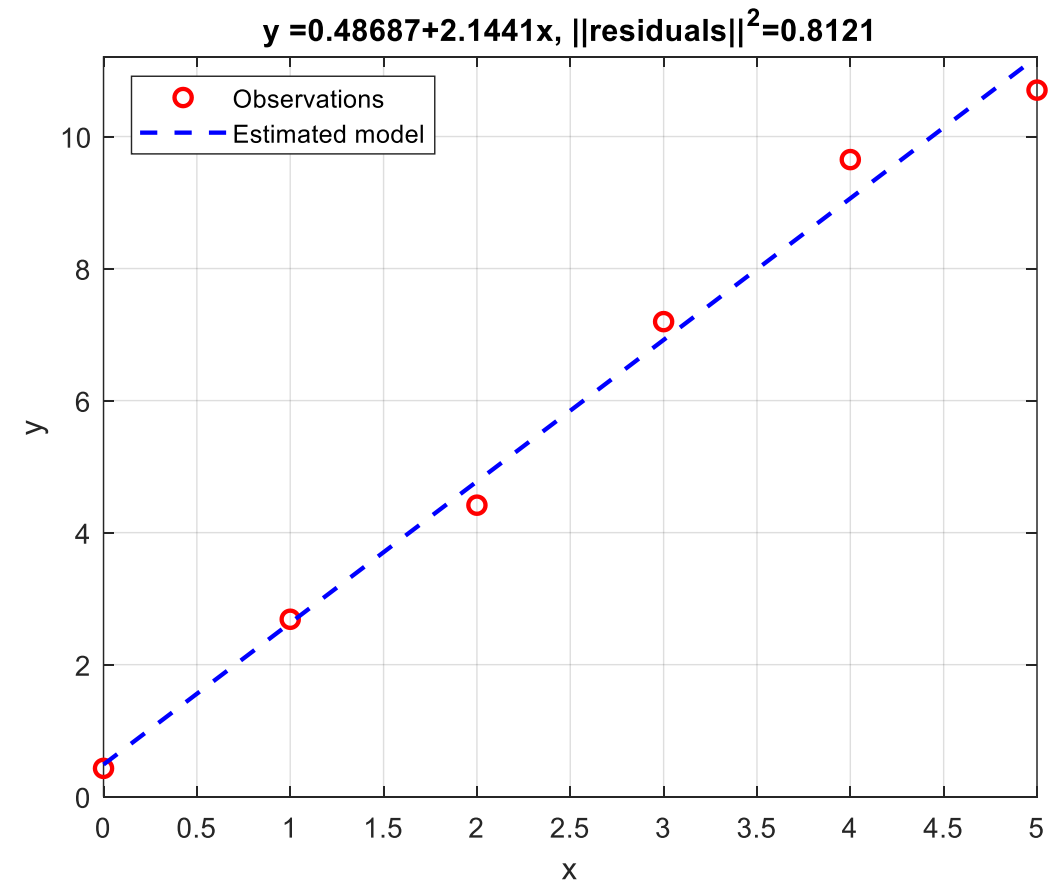
# Regresión lineal: Ejemplo 1, ajuste lineal

## □ Ej. 1) Ajuste de una línea recta $y = mx + b$

- Problema: encontrar  $m$  y  $b$  tal que ajusten el siguiente set de datos usando mínimos cuadrados ordinarios:
- Planteamiento del problema:

x	y
0	0.4288
1	2.6876
2	4.4156
3	7.1963
4	9.6509
5	10.7032

$$\begin{bmatrix} 0.4288 \\ 2.6876 \\ \vdots \\ 10.7032 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 5 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix}$$





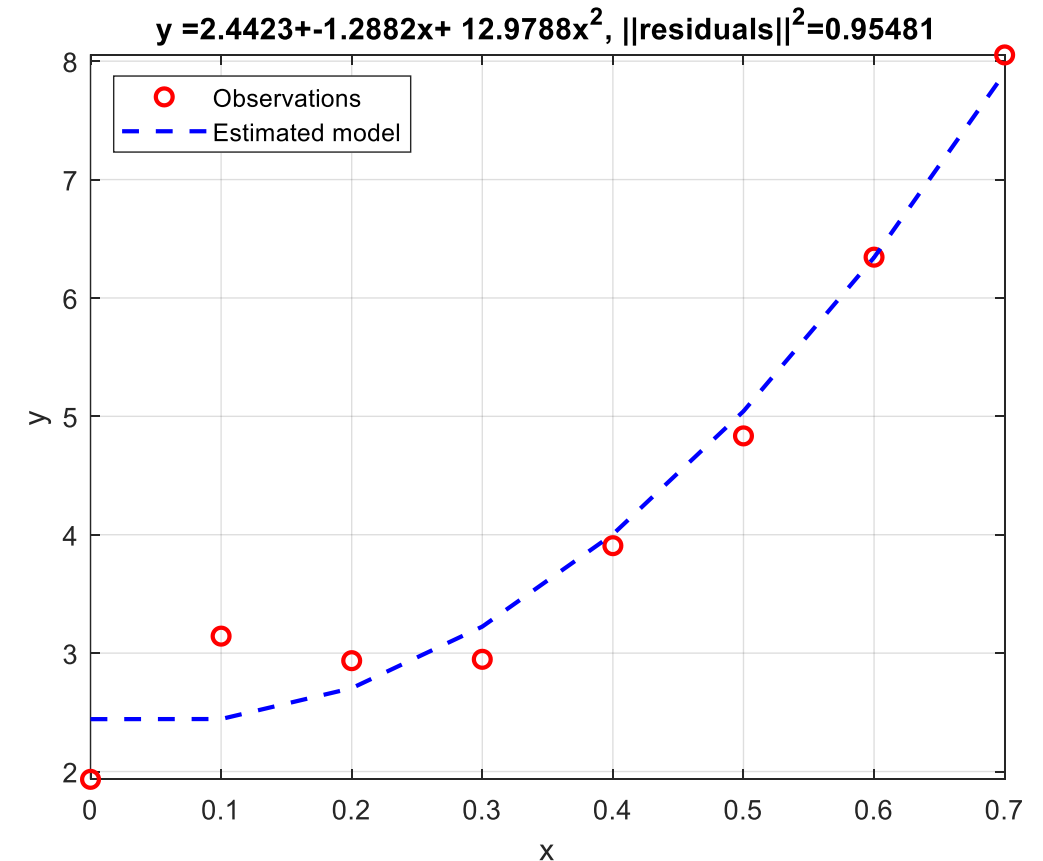
# Regresión lineal: Ejemplo 2, ajuste polinomial

□ **Ej. 2) Ajuste polinomial:**  $y = a_0 + a_1x + a_2x^2$

▪ Problema: encontrar  $a_0$ ,  $a_1$  y  $a_2$  para ajustar el siguiente conjunto de datos:

x	y
0	1.9339
0.1	3.1434
0.2	2.9363
0.3	2.9478
0.4	3.9076
0.5	4.8349
0.6	6.3447
0.7	8.0528

$$\begin{bmatrix} 1.9339 \\ 3.1434 \\ \vdots \\ 8.0528 \end{bmatrix} = \begin{bmatrix} 1 & x_i & x_i^2 \\ 1 & 0.1 & 0.1^2 \\ \vdots & \vdots & \vdots \\ 1 & 0.7 & 0.7^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

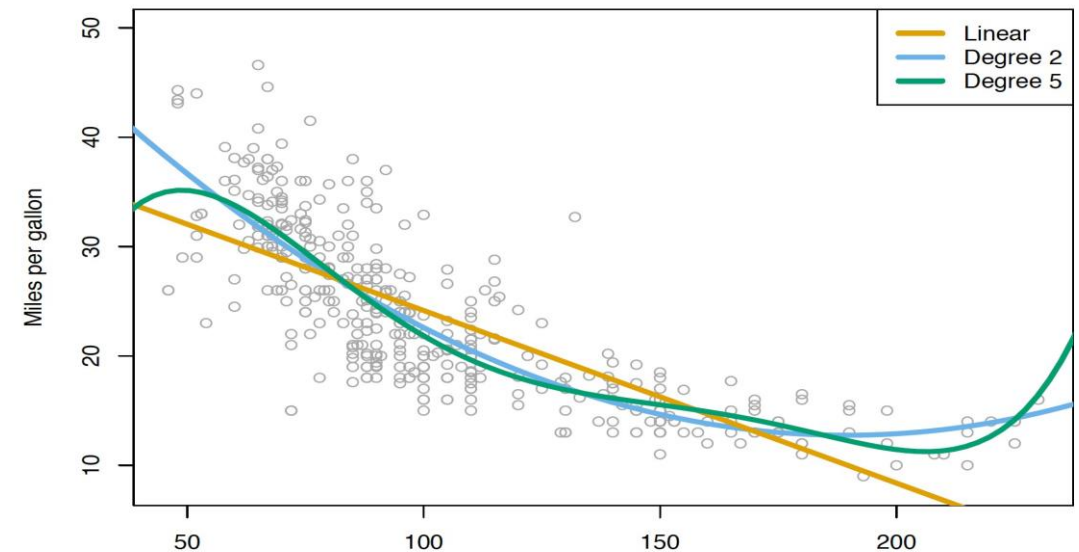


# Regresión lineal: Ejemplo 2, ajuste polinomial

- La regresión polinómica es una extensión de la regresión lineal. En general, la complejidad de un conjunto de datos excede la posibilidad de que se pueda ajustar por una línea recta. Es decir, la falta de adecuación obvia ocurre si se utiliza el modelo de regresión lineal original. La solución es utilizar la regresión polinómica.

$$h_w(x) = w_1x + w_2x^2 + \dots + w_nx^n + b$$

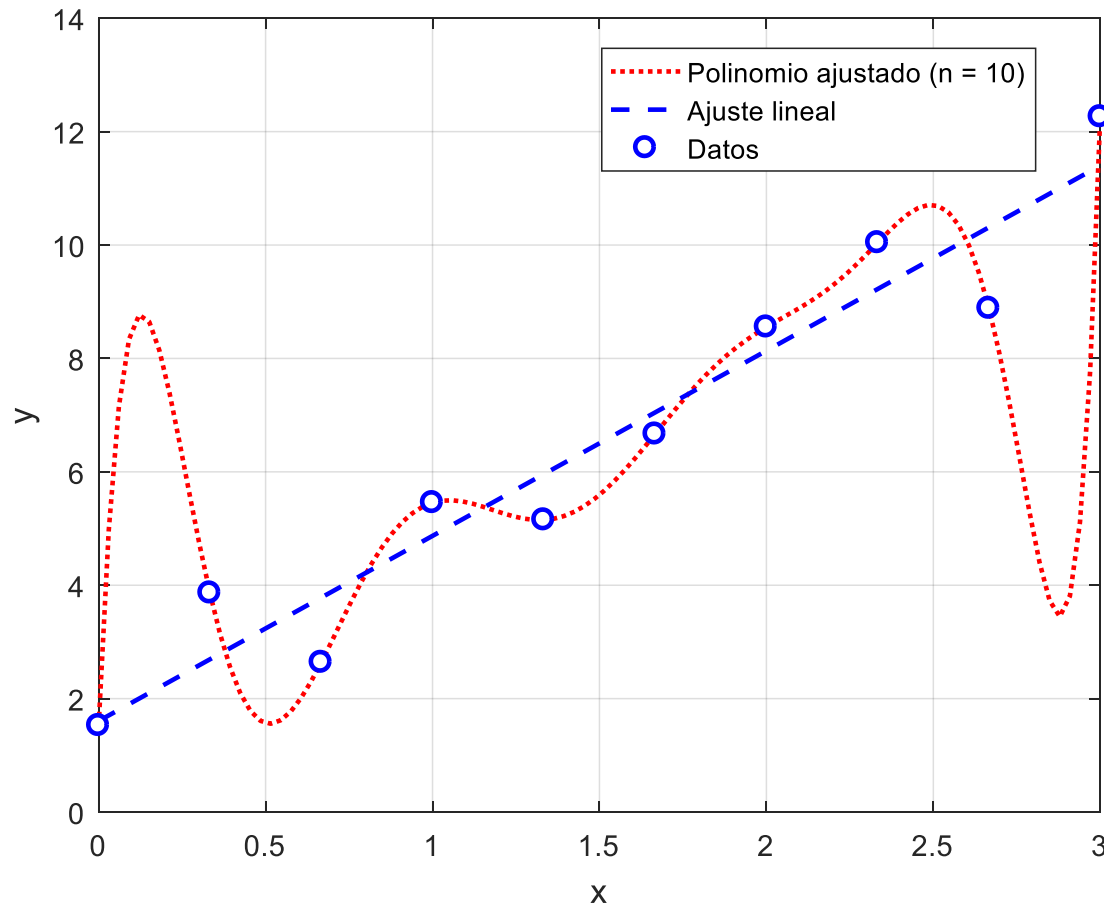
- donde, la n-ésima potencia es una dimensión de regresión polinomial (grado).
- La regresión polinomial pertenece a la regresión lineal en el sentido de que la relación entre sus parámetros  $w$  y la variable independiente sigue siendo lineal, mientras que su no linealidad se refleja en la dimensión de la característica.



Comparación entre regresión lineal y regresión polinómica

# El problema del sobreajuste

En la siguiente figura se ajustaron los datos a un polinomio de grado 10, evidentemente, los datos siguen una tendencia lineal en el rango observado, por lo que es suficiente con ajustar una línea recta a los datos.



## Recomendaciones:

- i) Conocer el problema y procedencia de los datos en detalle junto con las restricciones que los acompañan;
- ii) Usar estrategias de regularización;
- iii) Tener precaución al evaluar las métricas de calidad de ajuste, puede inducir a una mala selección, en el ejemplo, claramente el valor RMSE para el caso del polinomio es 0 en los datos ajustados.

# El problema del sobreajuste: Prevención usando regularización

- ❑ Los términos de **regularización (regularization)** se pueden utilizar para reducir el sobreajuste. El valor de  $w$  no puede ser demasiado grande o demasiado pequeño en el espacio de muestra. Para evitar esto, se puede agregar una función de pérdida en la función de destino basada en algunas normas.
- ❑ **Términos de regularización (norma):** Por ejemplo, el término de regularización se puede implementar con la norma-L2. La regresión lineal que utiliza esta función de pérdida también se denomina regresión **Ridge**.
- ❑ La regresión lineal con pérdida absoluta se denomina regresión de **Lasso**.

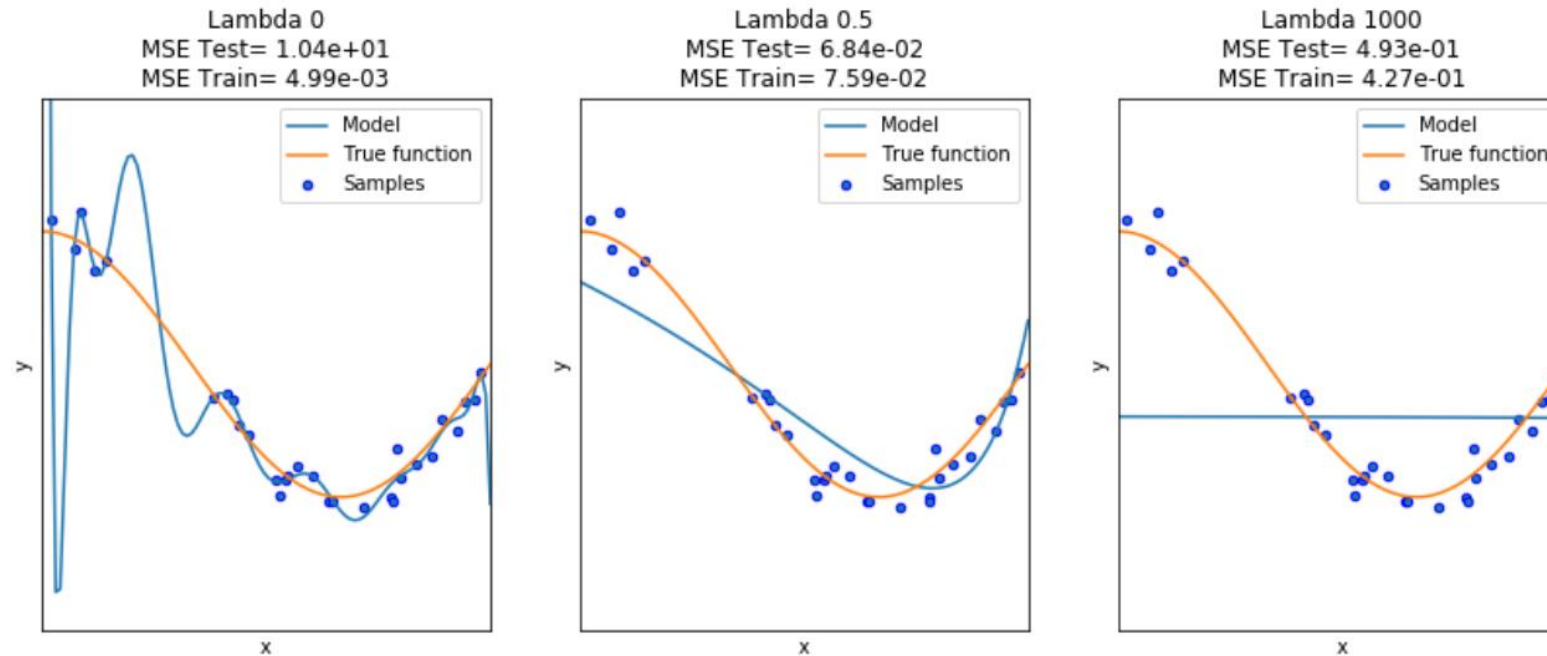
$$J(w) = \frac{1}{2m} \sum (h_w(x) - y)^2 + \lambda \sum \|w\|_2^2$$

$$J(w) = \frac{1}{2m} \sum (h_w(x) - y)^2 + \lambda \sum \|w\|_1$$

$\lambda$  es un hiperparámetro que indica con que fuerza se penalizarán los parámetros, mientras mayor sea su valor, los parámetros tenderán más rápido a cero durante el entrenamiento.

# El problema del sobreajuste: Prevención usando regularización

□ Efecto del parámetro de regularización en una regresión:



Ajuste con polinomio de grado 15 y sin regularización.

Efecto de regularización al variar parámetro lambda.

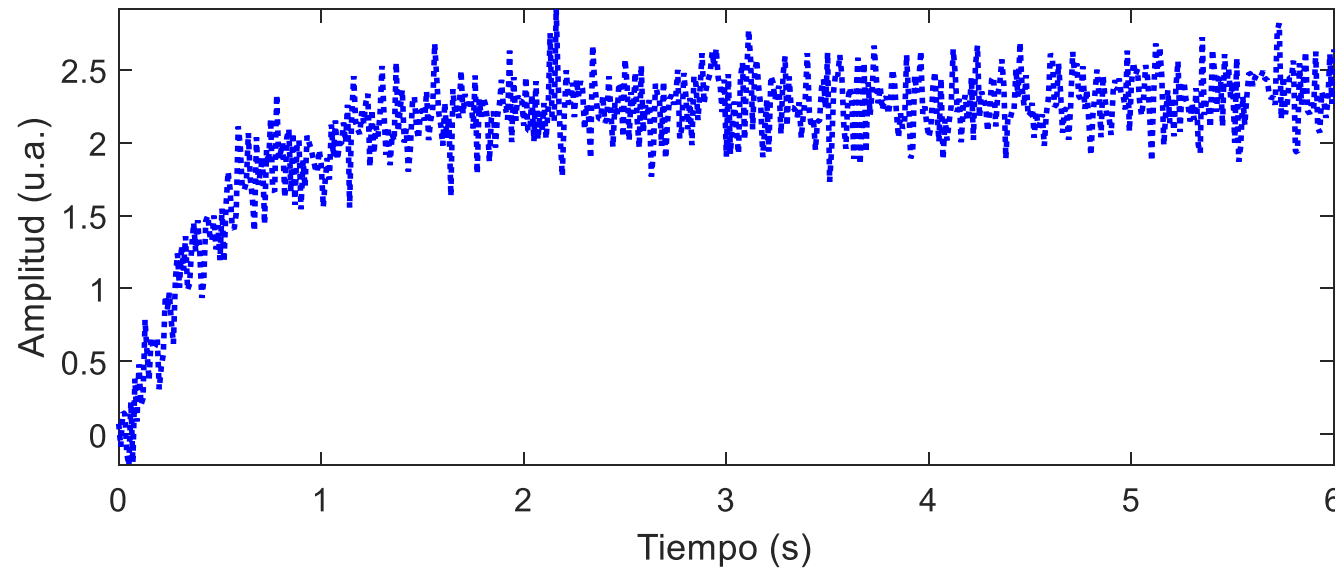
[fuente](#)

# Ajuste de curvas no lineales

Este tipo de problemas también se puede enfrentar cómo los anteriores mediante un proceso de optimización, el detalle es que ahora los parámetros a optimizar no estarán combinados linealmente con los datos ni transformaciones de los datos. Los parámetros aparecerán dentro de funciones no lineales, como el ejemplo de más abajo.

**Ejemplo:** Se midió la respuesta de un sistema a una entrada tipo escalón unitario como mostrada en la figura, se pide encontrar los parámetros relevantes del sistema si se sabe que el modelo tiene la forma:

$$y(t) = k(1 - e^{-t/\tau}), t \geq 0$$



**Qué parámetros queremos estimar?**

- Este y otros problemas los resolveremos en la práctica de regresión



# Caso Real

- ❑ Predicción de la ley de Cobre en muestras de metal blanco

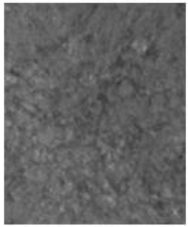
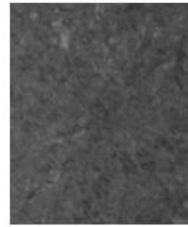
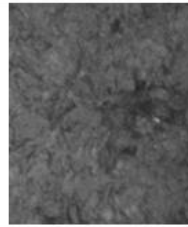
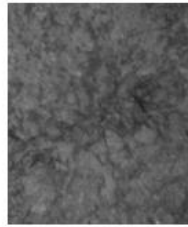
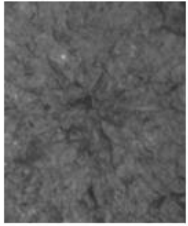
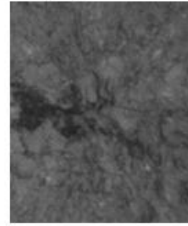
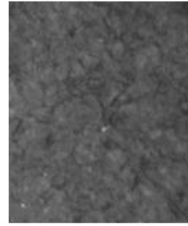
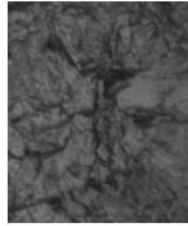
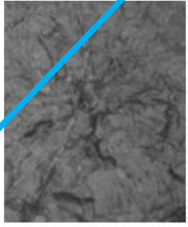
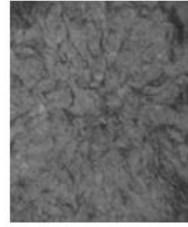
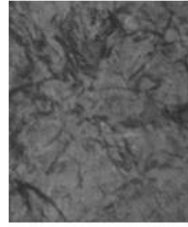
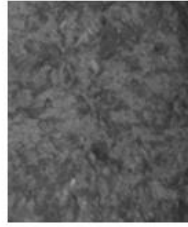
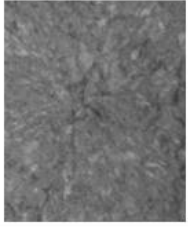
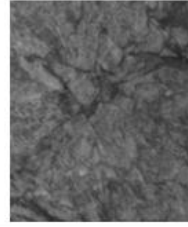
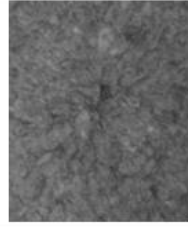
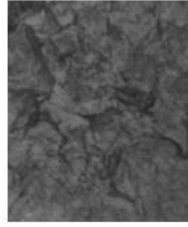


# Caso Real

❑ Para un conjunto de imágenes de muestras de metal blanco (uno de los productos en el proceso de extracción de cobre), se tienen las concentraciones de cobre respectivas obtenidas mediante métodos analíticos.

❑ El objetivo fue generar modelos de regresión lineal para estimar la concentración de cobre (variable dependiente) cuando se tuvieran nuevas imágenes de la muestra.

❑ Las variables independientes corresponden a características extraídas desde las imágenes.

			
Sample_1: 56.7%Cu	Sample_2: 57.1%Cu	Sample_3: 58.0%Cu	Sample_4: 60.2%Cu
			
Sample_5: 60.2%Cu	Sample_6: 60.5%Cu	Sample_7: 60.5%Cu	Sample_8: 64.5%Cu
			
Sample_9: 64.1%Cu	Sample_10: 62.5%Cu	Sample_11: 63.0%Cu	Sample_12: 63.0%Cu
			
Sample_13: 60.9%Cu	Sample_14: 60.9%Cu	Sample_15: 59.4%Cu	Sample_16: 62.3%Cu

$$y = a_0 + a_1x_1 + a_2x_2$$