Modelos de datos en Python

Juan Fernando Pérez
Departamento de Ingeniería Industrial
Universidad de los Andes

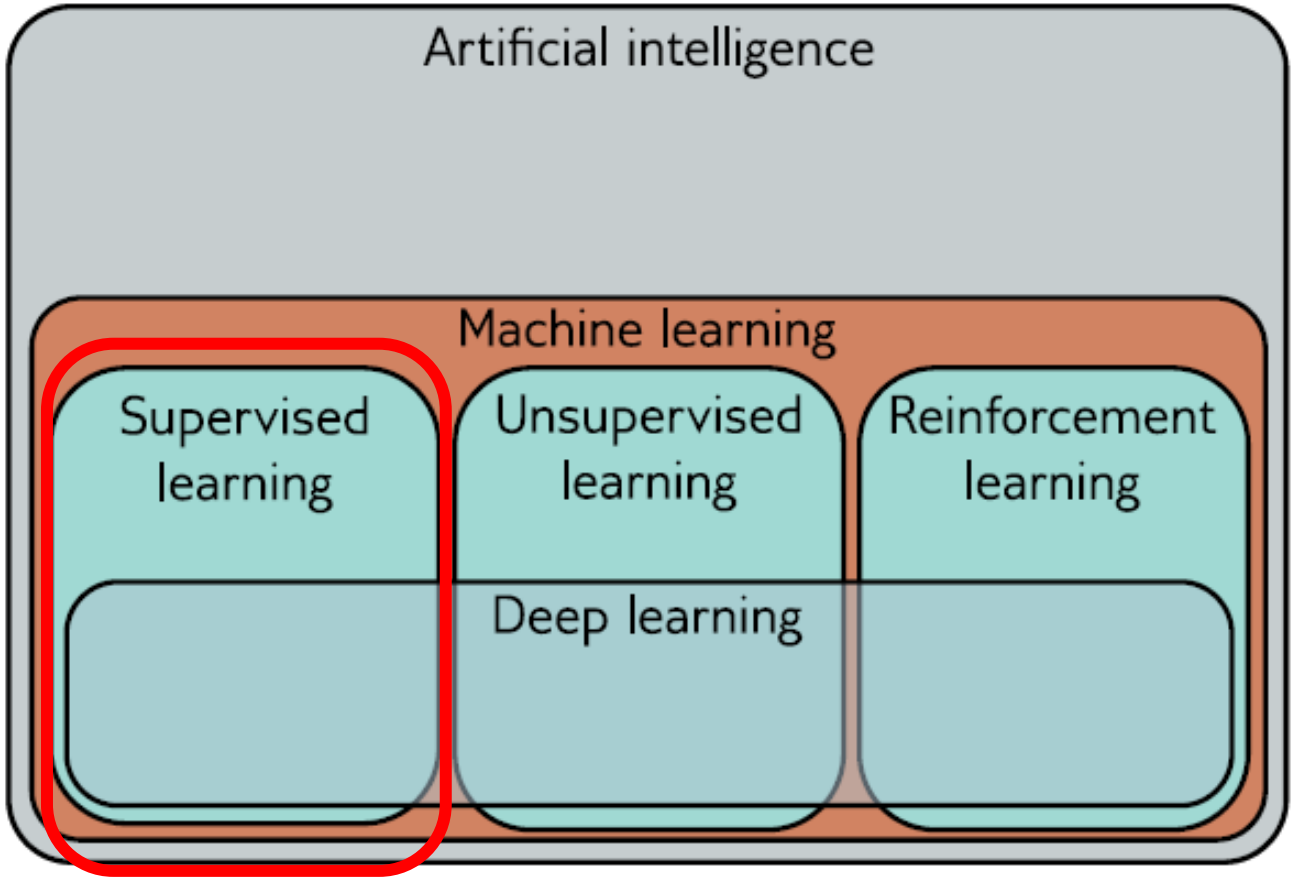
Agosto de 2024

jf.perez33@uniandes.edu.co
juanfperez.com

Universidad de los Andes

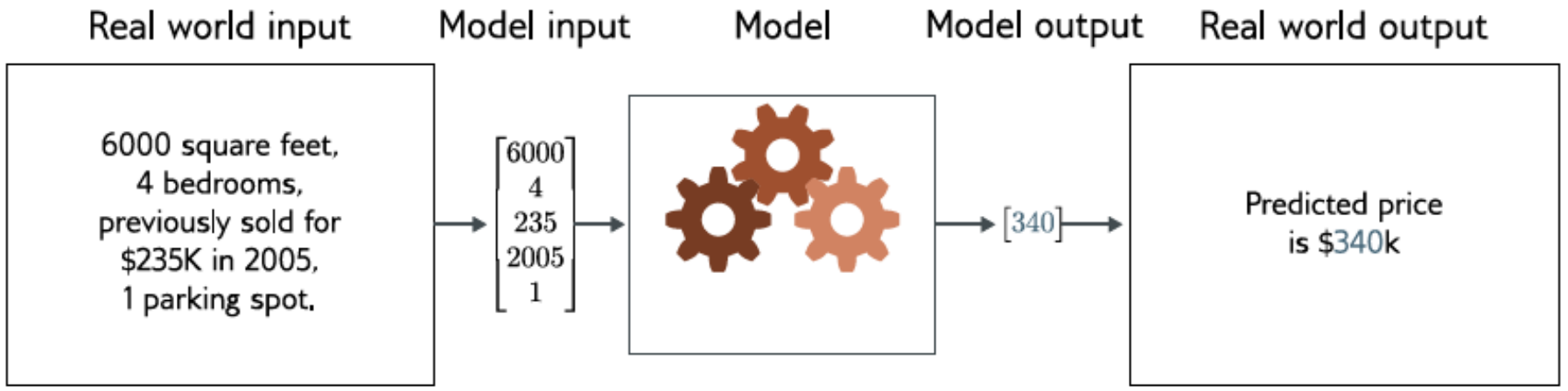
Aprendizaje supervisado

Problemas de Aprendizaje de máquina



Crédito: Prince, 2023

Aprendizaje supervisado

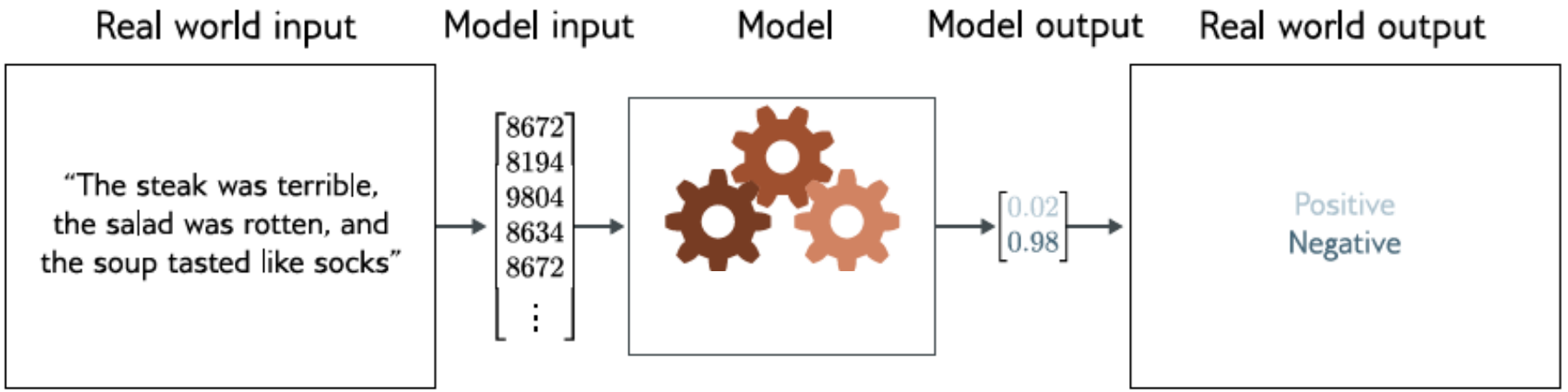


Predicción de una **salida continua**: valor de un inmueble

Regresión

Crédito: Prince, 2023

Aprendizaje supervisado

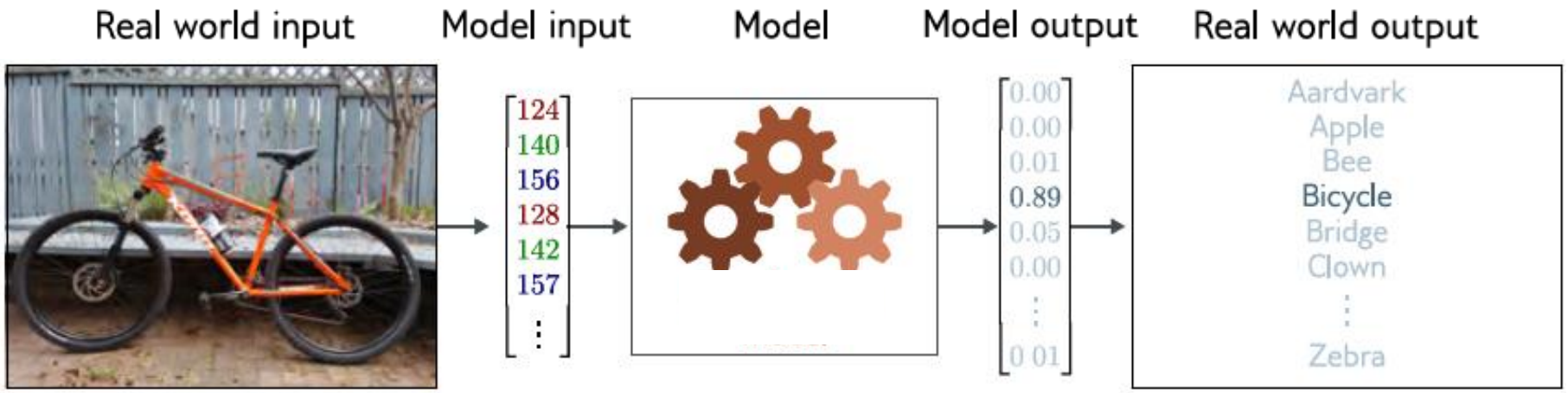


Predicción de una **salida discreta** (categoría, **binaria**)

Clasificación

Crédito: Prince, 2023

Aprendizaje supervisado

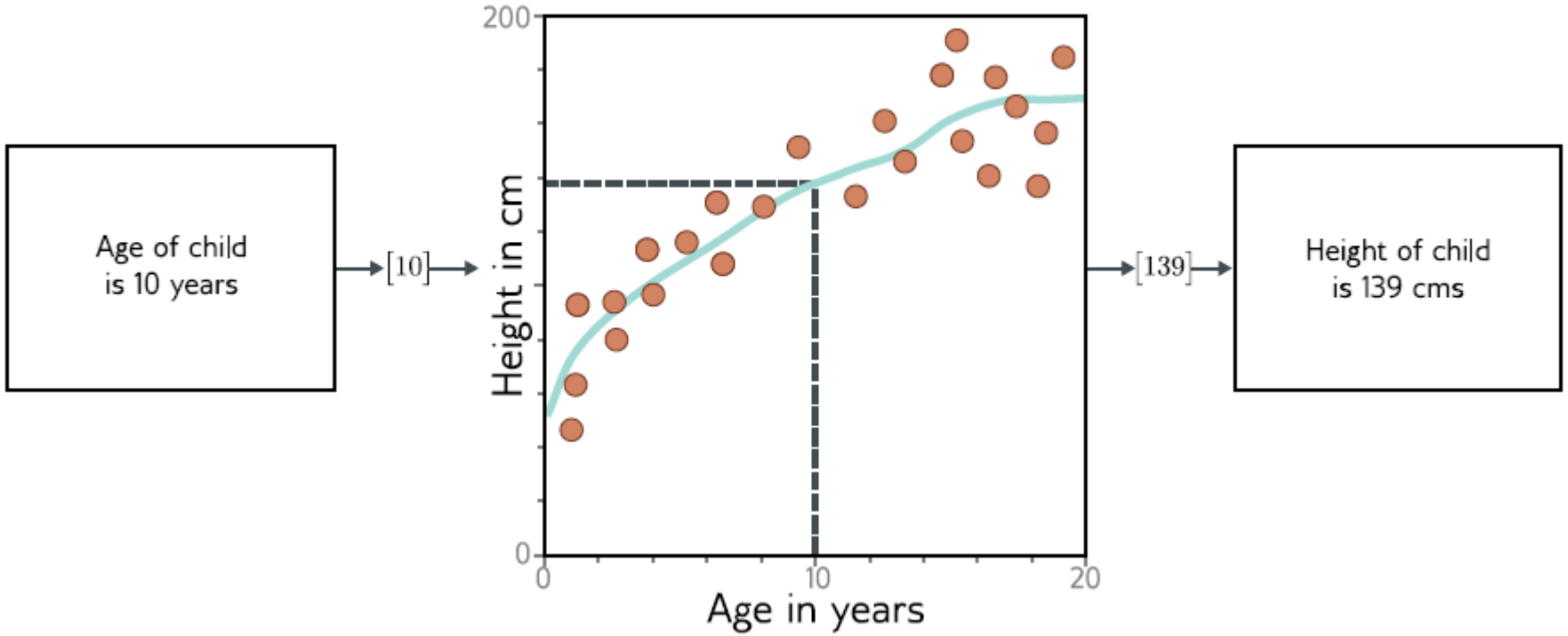


Predicción de una **salida discreta** (categoría, **multiclase**)

Clasificación

Crédito: Prince, 2023

Aprendizaje supervisado – Ej. Regresión



Modelo de aprendizaje de máquina: **familia** de **ecuaciones** que permiten obtener una **salida** a partir **de una entrada**

Crédito: Prince, 2023

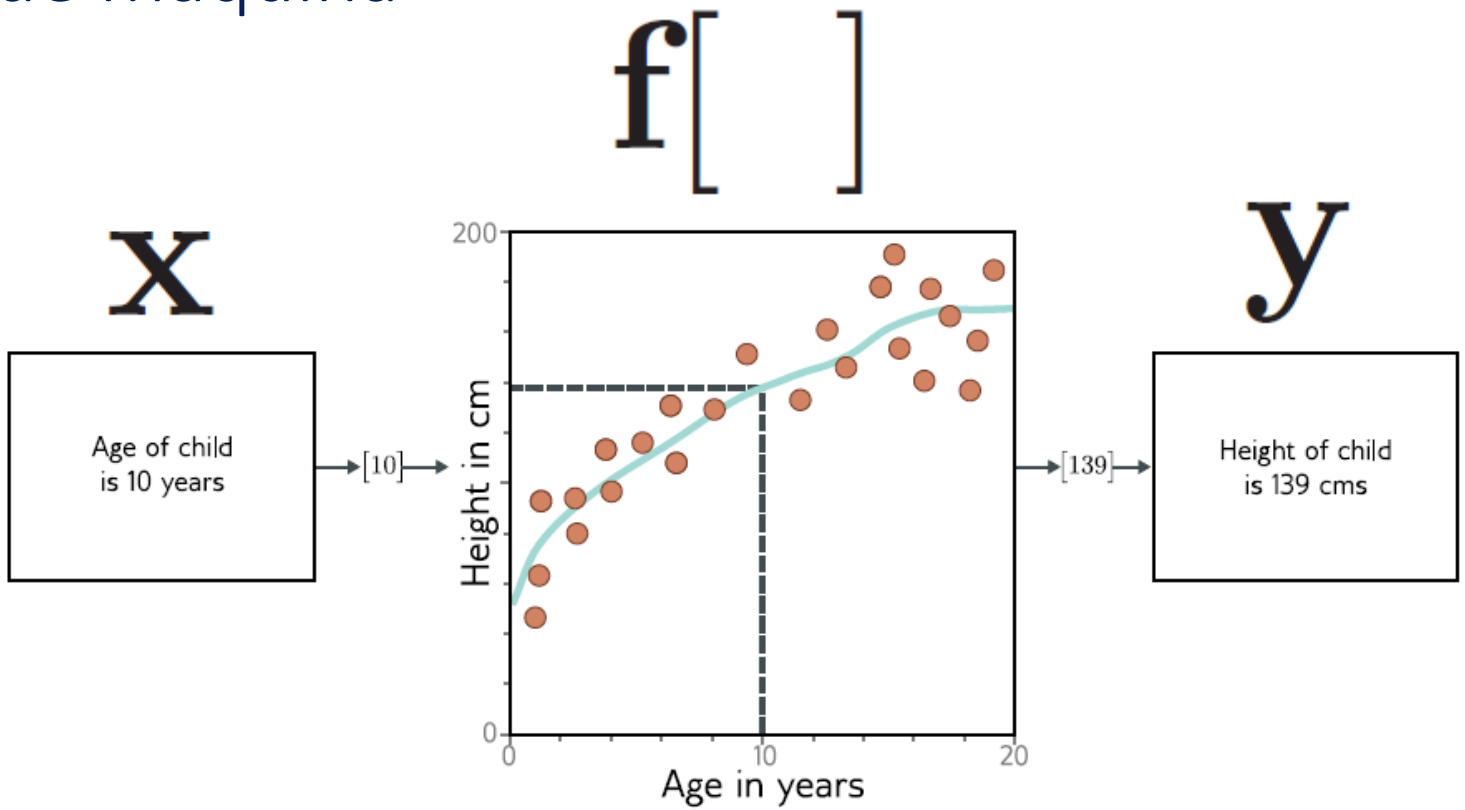
Aprendizaje supervisado

$$\mathbf{y} = \mathbf{f}[\mathbf{x}]$$

f: Modelo de aprendizaje de máquina

x: entrada

y: salida



Crédito: Prince, 2023

Aprendizaje supervisado

$$\mathbf{y} = \mathbf{f}[\mathbf{x}, \boldsymbol{\phi}]$$

f: Modelo de aprendizaje de máquina

x: entrada

y: salida

ϕ: parámetros que definen el modelo

Entrenamiento

Determinar el **mejor valor** de los **parámetros** considerando un **conjunto** de **datos** (de entrenamiento)

Aprendizaje supervisado - Entrenamiento

$$\mathbf{y} = \mathbf{f}[\mathbf{x}, \boldsymbol{\phi}]$$

Datos de entrenamiento: $\{\mathbf{x}_i, \mathbf{y}_i\}$

Observaciones indexadas por i

x: entrada observada

y: salida observada (etiqueta)

Supervisado

Se requieren **ejemplos** pasados de **parejas entrada-salida** para entrenar el modelo

Aprendizaje supervisado - Entrenamiento

¿Cómo cuantificar qué tan bueno es un modelo $\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}]$?

Función de pérdida (loss): $\mathrm{L}[\boldsymbol{\phi}]$

Valores **bajos** si con los parámetros el modelo captura **bien** la relación entre las entradas y salidas de los datos de entrenamiento

Aprendizaje supervisado - Entrenamiento

¿Cómo cuantificar qué tan bueno es un modelo $\mathbf{f}[\mathbf{x}, \boldsymbol{\phi}]$?

Función de pérdida (loss): $\mathrm{L}[\boldsymbol{\phi}]$

Mejor valor de los parámetros: $\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[\mathrm{L}\left[\boldsymbol{\phi}\right]\right]$

*$\mathrm{L}\left[\{\mathbf{x}_i, \mathbf{y}_i\}, \boldsymbol{\phi}\right]$: depende de los datos de entrenamiento

Aprendizaje supervisado: regresión

Entrenamiento: ejemplo Regresión Simple

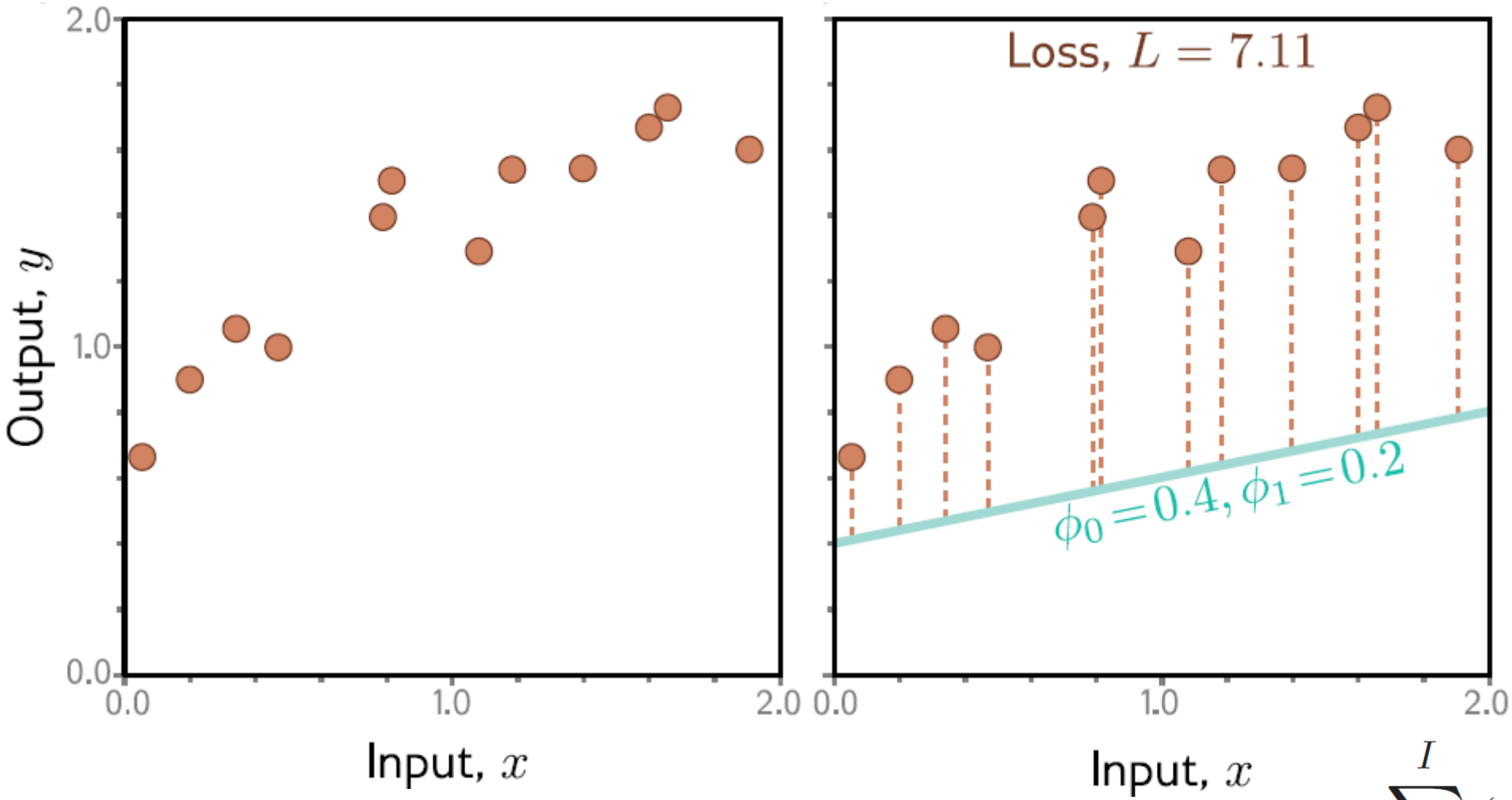
Modelo: $$y = \mathrm{f}[x, \boldsymbol{\phi}] = \phi_0 + \phi_1 x$$

Función de pérdida:

$$\mathrm{L}[\boldsymbol{\phi}] = \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2$$

$$= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2$$

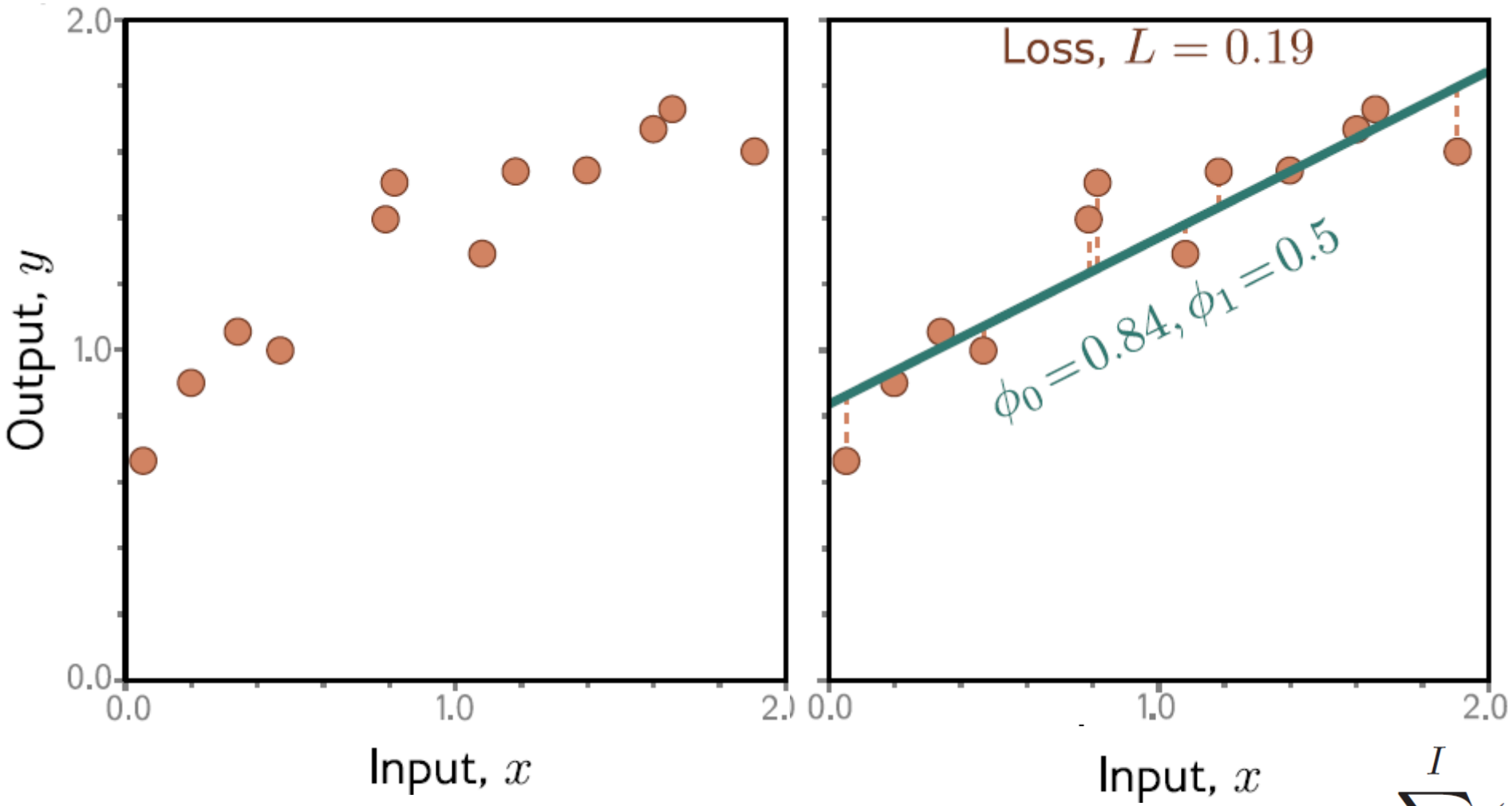
Entrenamiento: ejemplo Regresión Simple



$$\sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2$$

Crédito: Prince, 2023

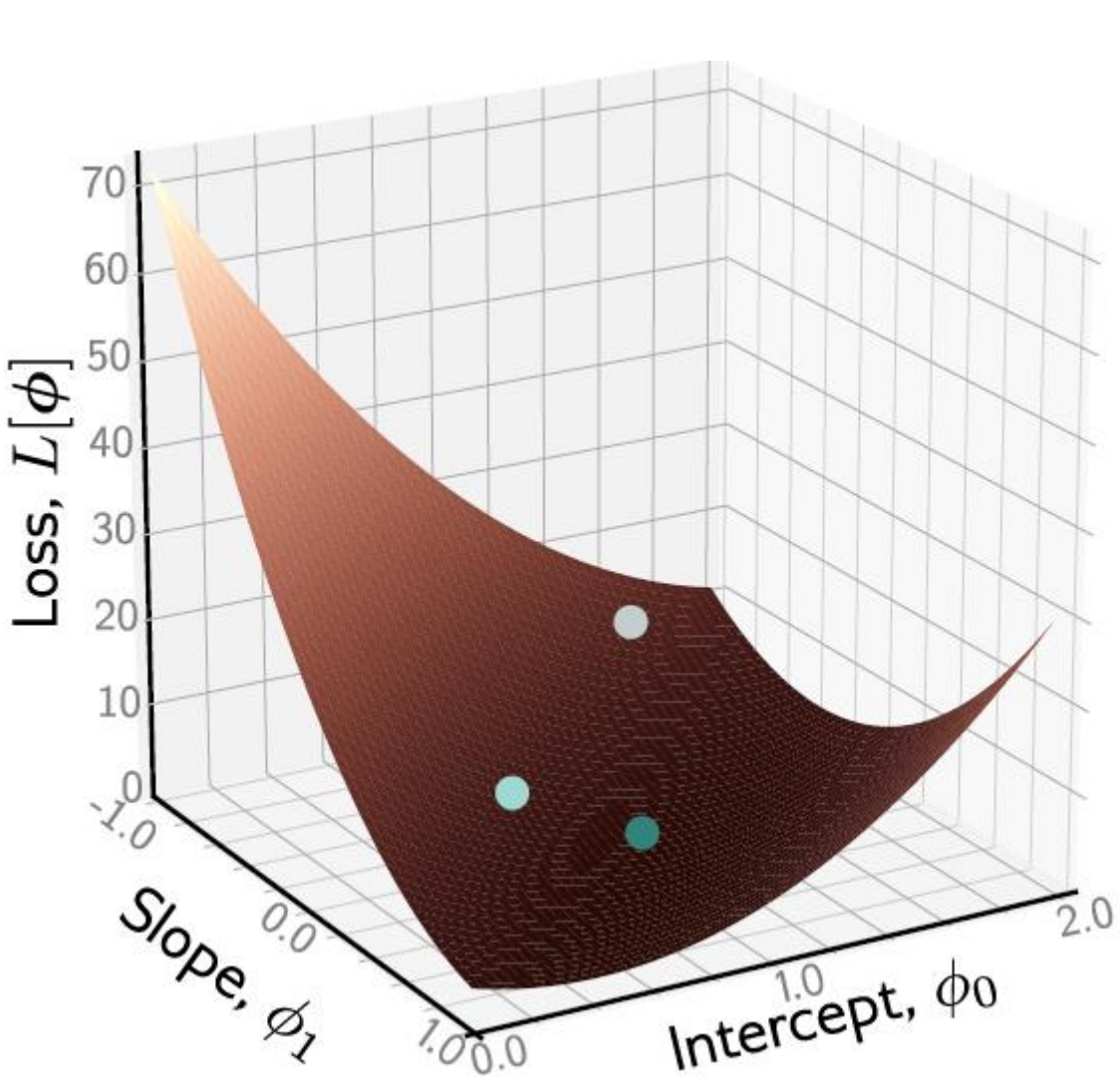
Entrenamiento: ejemplo Regresión Simple

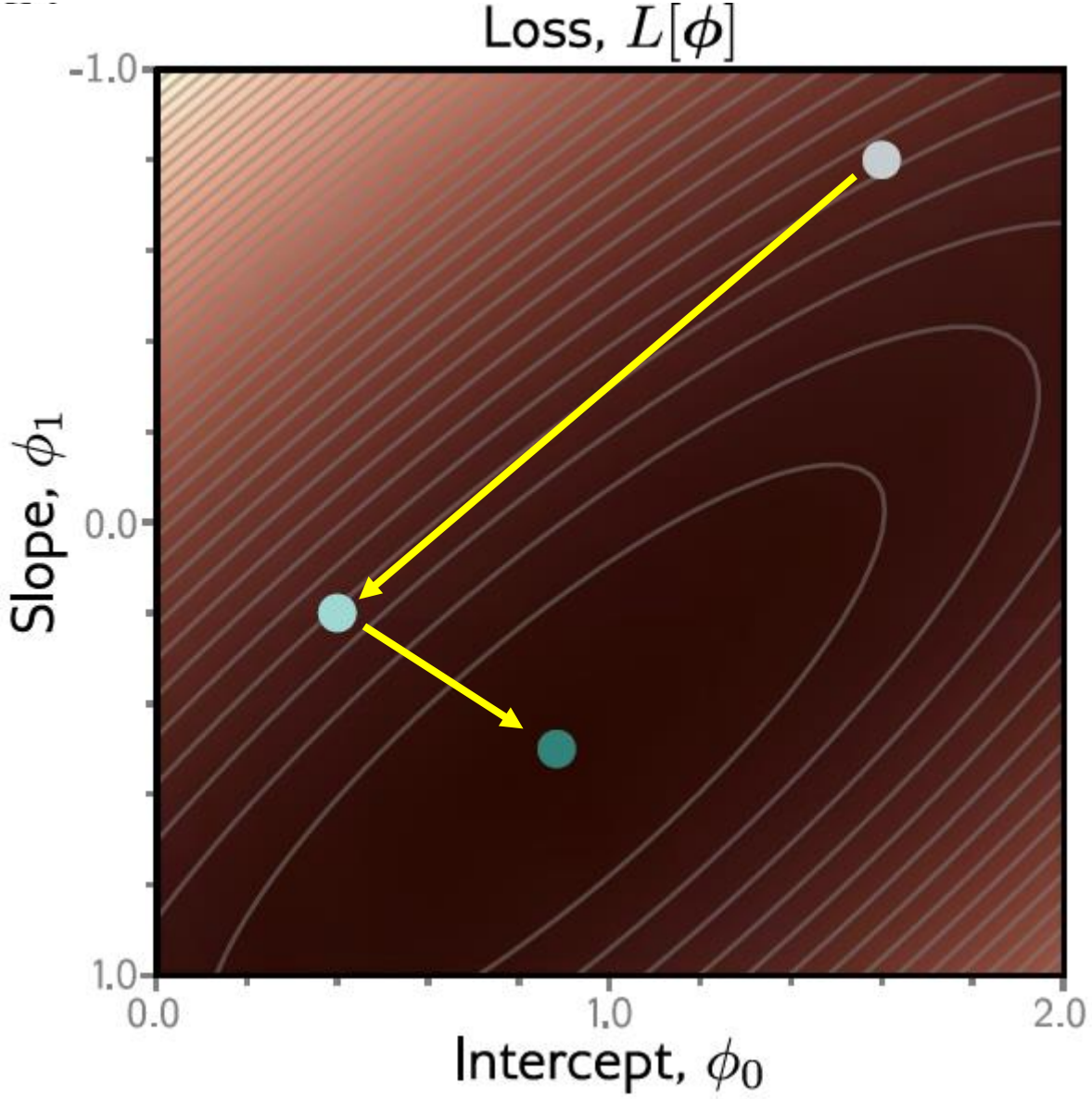


$$\sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2$$

Crédito: Prince, 2023

Entrenamiento: ejemplo Regresión Simple

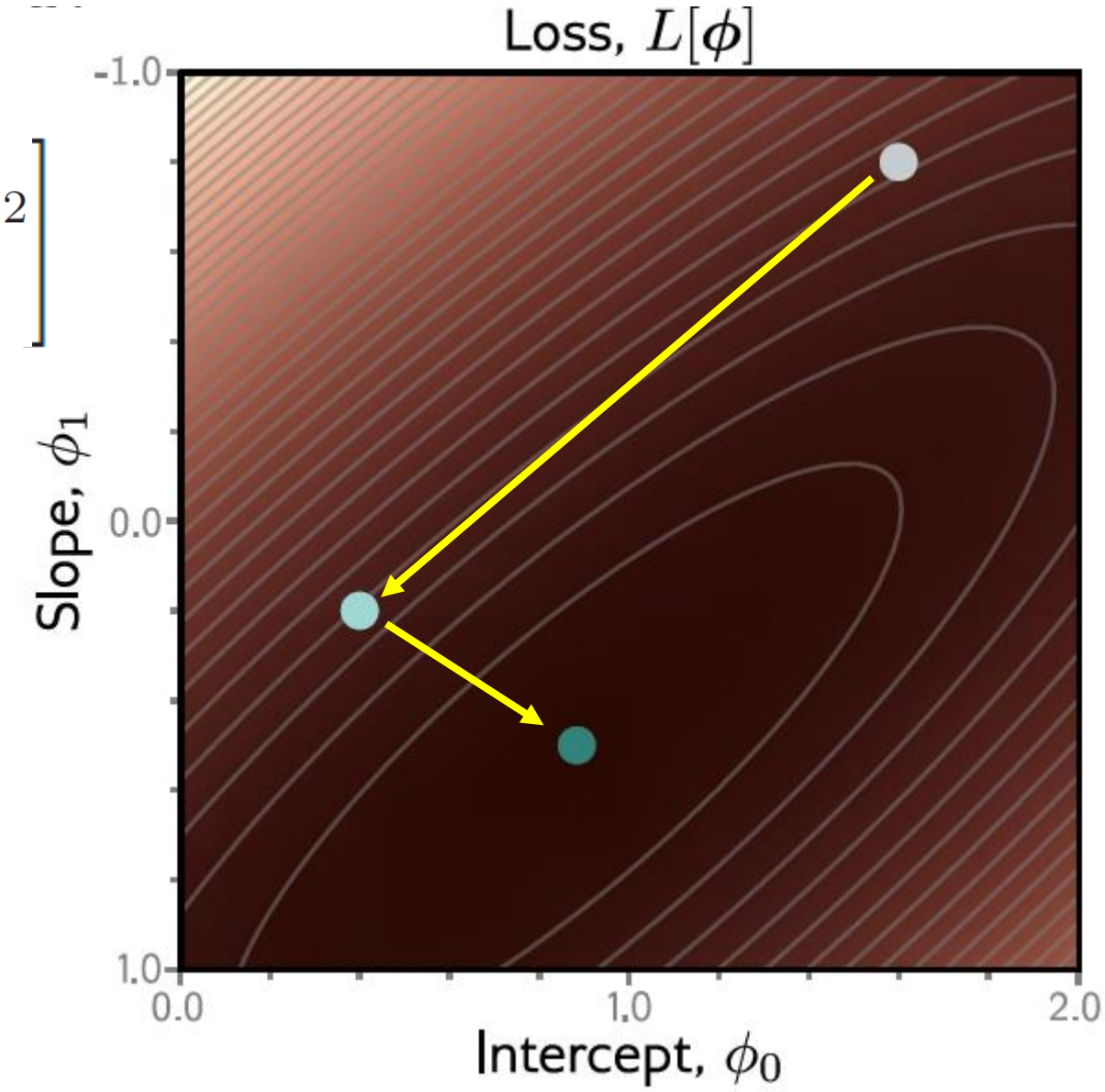




Crédito: Prince, 2023

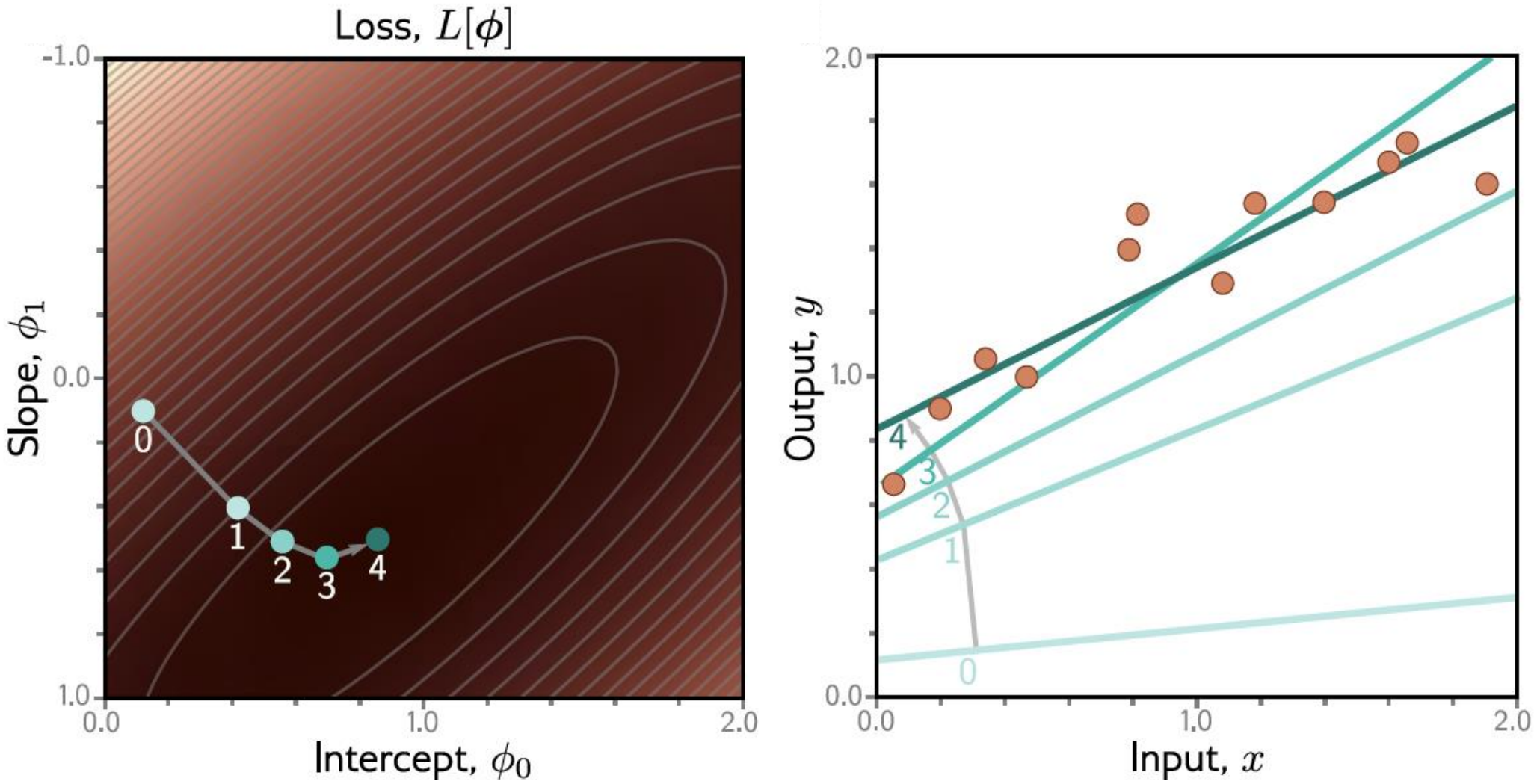
Entrenamiento: ejemplo Regresión Simple

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left[\sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2 \right]$$



Crédito: Prince, 2023

Entrenamiento: ejemplo Regresión Simple



Crédito: Prince, 2023

Entrenamiento, sobreajuste y subajuste

Aprendizaje supervisado - Entrenamiento

Entrenamiento con datos (de entrenamiento)

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}} \Big[\mathrm{L} \left[\phi \right] \Big]$$

Capacidad de **generalizar** a otros datos: **datos de prueba**

Datos

Aprendizaje supervisado - Entrenamiento

Entrenamiento con datos (de entrenamiento)

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}} \Big[\mathrm{L} \left[\phi \right] \Big]$$

Capacidad de **generalizar** a otros datos: **datos de prueba**

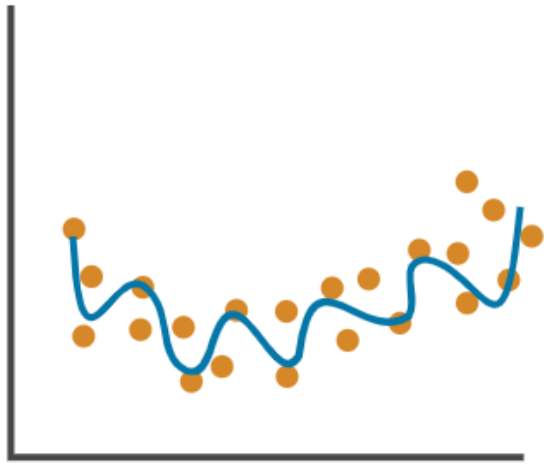
Datos de entrenamiento

Datos de prueba

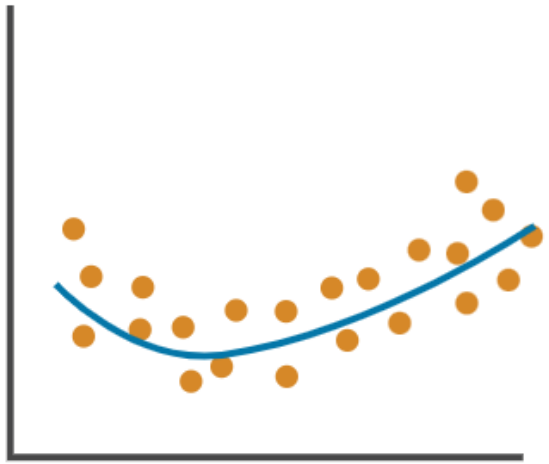
Sobreajuste vs subajuste

Aprendizaje supervisado - Entrenamiento

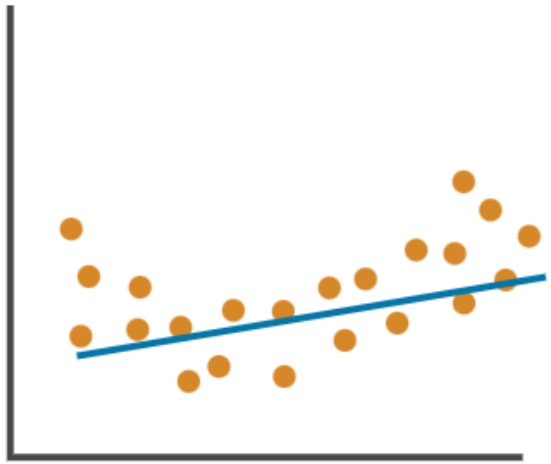
Sobreajuste



Ajuste



Subajuste



Crédito: AWS, 2024

Validación cruzada

Usar **todos** los datos para entrenamiento.

Validación cruzada en k grupos (k-fold cross validation).

Realizar k veces la separación, con subconjuntos diferentes.

Universidad de los Andes

Validación cruzada

Usar **todos** los datos para entrenamiento.

Validación cruzada en k grupos (k-fold cross validation).

Realizar k veces la separación, con subconjuntos diferentes.

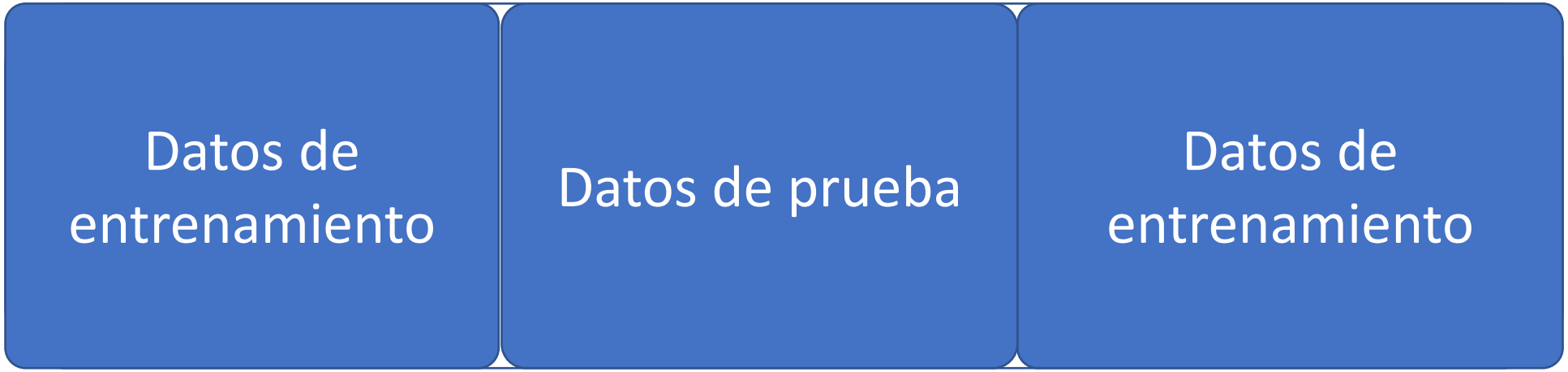
Universidad de los Andes

Validación cruzada

Usar **todos** los datos para entrenamiento.

Validación cruzada en k grupos (k-fold cross validation).

Realizar k veces la separación, con subconjuntos diferentes.



Universidad de los Andes

Aprendizaje como Optimización

Medida de **desempeño (loss)**: función **objetivo**

Espacio de búsqueda:

- **Variables** de **decisión: parámetros del modelo**
- **Restricciones sobre los parámetros**

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \Big[\mathrm{L} \left[\boldsymbol{\phi} \right] \Big]$$

Modelos de regresión en Python (Scikit Learn)

Regresión en Python – Scikit learn

https://scikit-learn.org/

Aprendizaje supervisado:

https://scikit-learn.org/stable/supervised_learning.html

fit() + predict()

Regresión en Python – Scikit learn

Train test split:

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Regresión en Python – Scikit learn

Cross validation:

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score

Universidad de los Andes

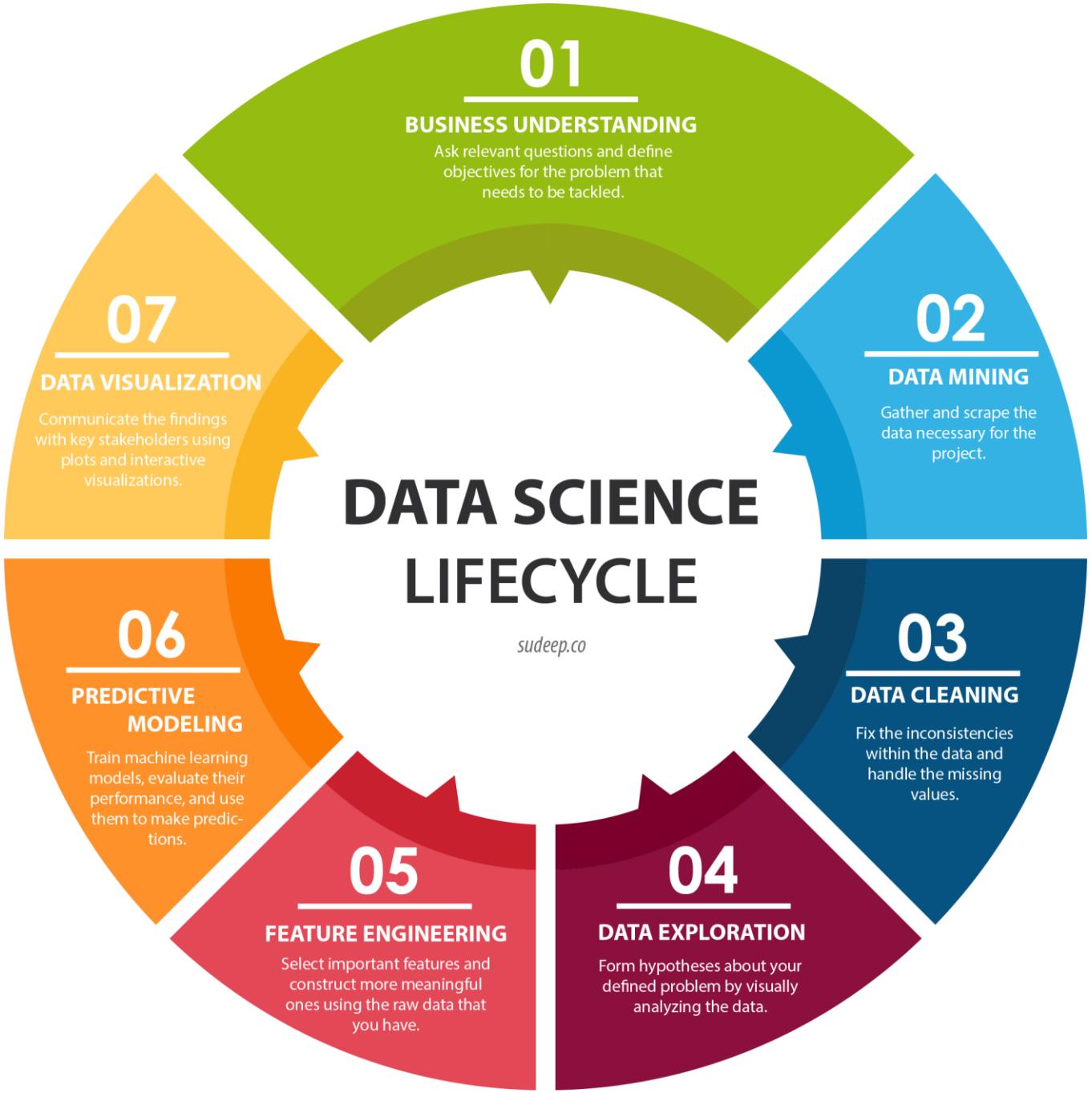
Regresión en Python – Statsmodels

API para regresión con OLS:

https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

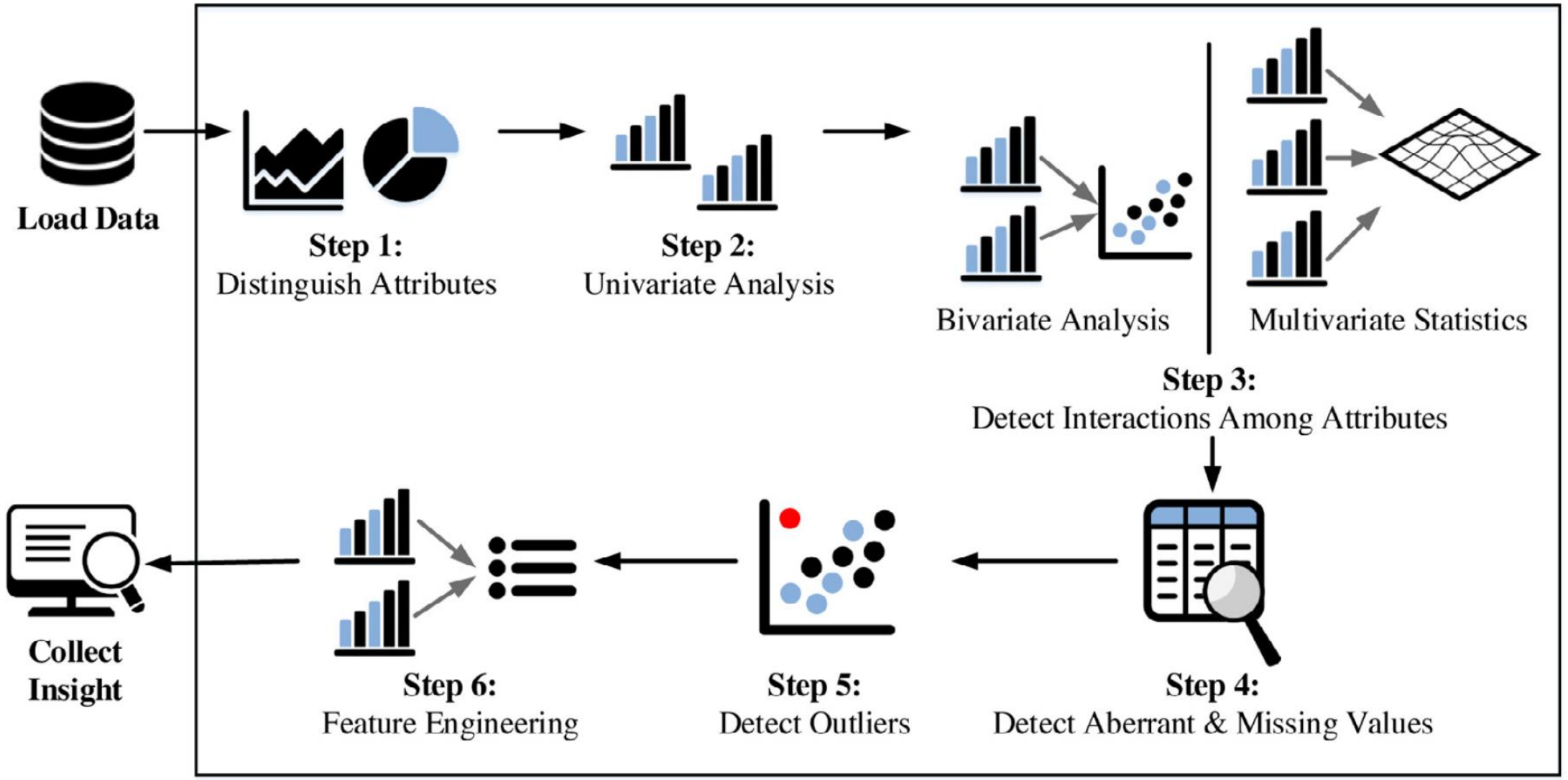
Metodologías para Análisis Exploratorio de Datos

Análisis Exploratorio (EDA) en Ciencia de Datos



Crédito: sudeep.co

Pasos de EDA



Crédito: A. Gosh (2018) A comprehensive review of tools for exploratory analysis of tabular industrial datasets

Lectura

A. Gosh (2018)

A comprehensive review of tools for exploratory analysis of tabular industrial datasets.

Sección 3.