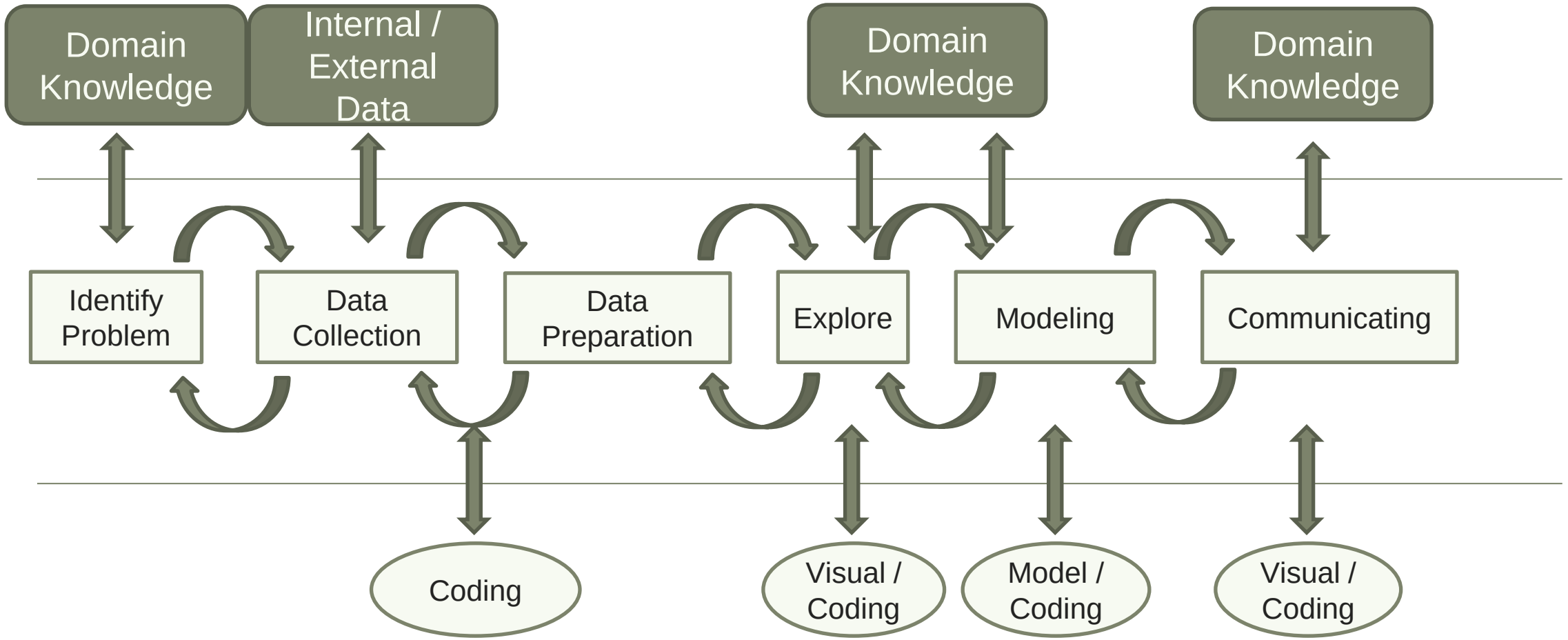
X1. What is data mining/business analytics?



- Generating actionable business insights from data
- An iterative process

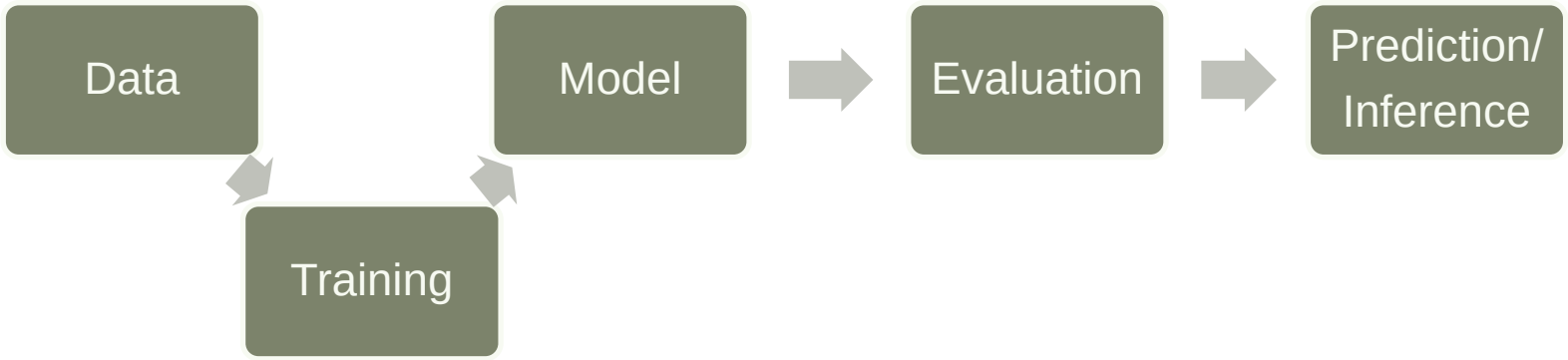
What is machine learning?

- “A field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.” (Wikipediae)
- “ML is the process of training a piece of software, called a model, to make useful predictions from data.” – Google AI

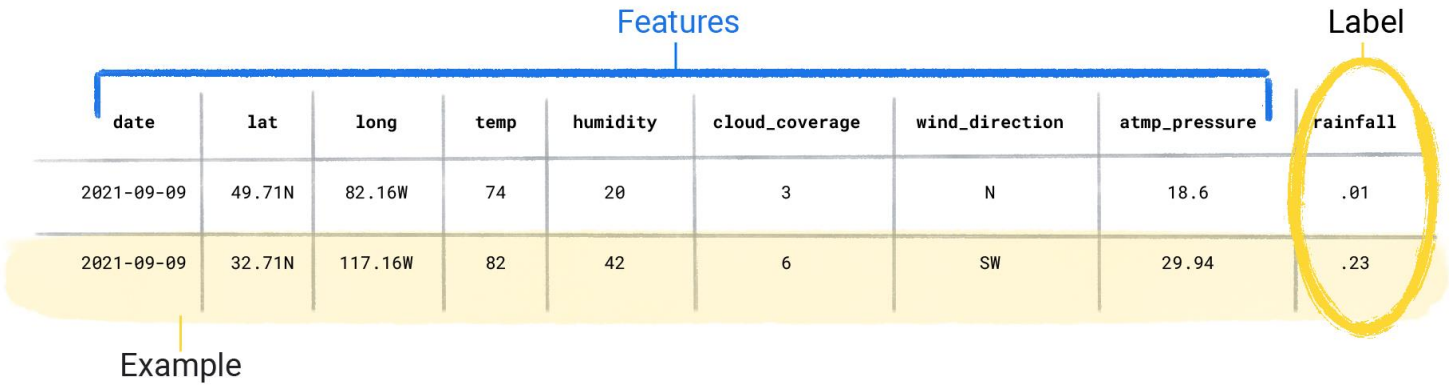
Supervised learning

- The correct answers (labels: Y) are given in the data
- Train a model to find the mapping (model) between features (Xs) and Y using data
- Analogy:
    - Students are given past questions and correct answers
    - Students learn by studying the questions and answers
    - Student can answer new questions based on training

Key components of Supervised Learning

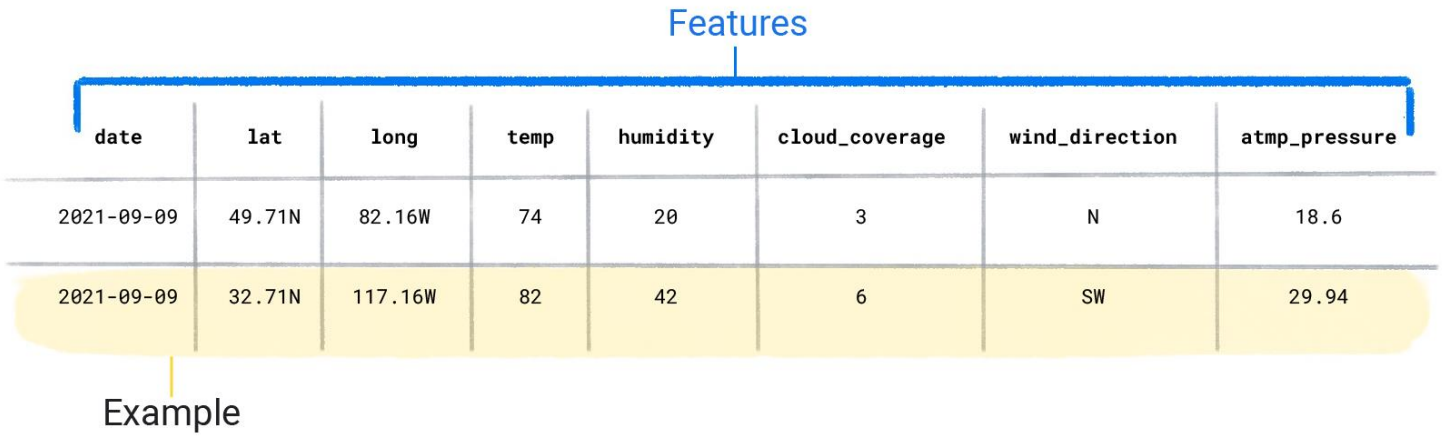


Data



Features: attributes, fields, independent variables, predictors, explanatory variables

Label: dependent variable, target, response



Data - Notation

In general, we will let x_{ij} represent the value of the jth variable for the ith observation, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. Throughout this book, i will be used to index the samples or observations (from 1 to n) and j will be used to index the variables (from 1 to p). We let $\mathbf{X}$ denote a $n \times p$ matrix whose (i, j)th element is x_{ij}. That is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

- You can visualize X as a spreadsheet of numbers with *n* rows and *p* columns

Data - Notation

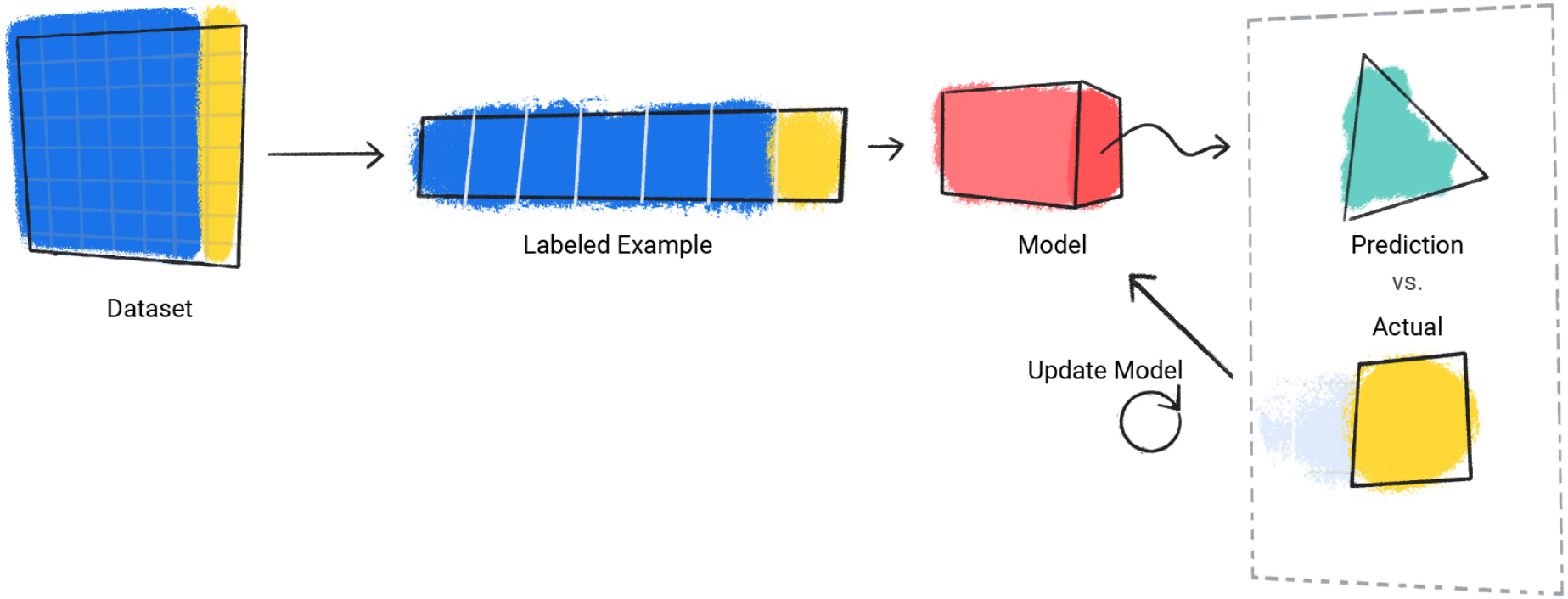
We use y_i to denote the ith observation of the variable on which we wish to make predictions, such as `wage`. Hence, we write the set of all n observations in vector form as

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Then our observed data consists of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each x_i is a vector of length p. (If $p = 1$, then x_i is simply a scalar.)

Training

- To find the best solution for predicting Y from Xs
- By comparing prediction with actual values, the model keeps updating itself
- Finally, it finds the best model to make predictions



Source: Google AI

Model

A mathematical mapping between features (Xs) to output labels (Y)
Most often seen models:

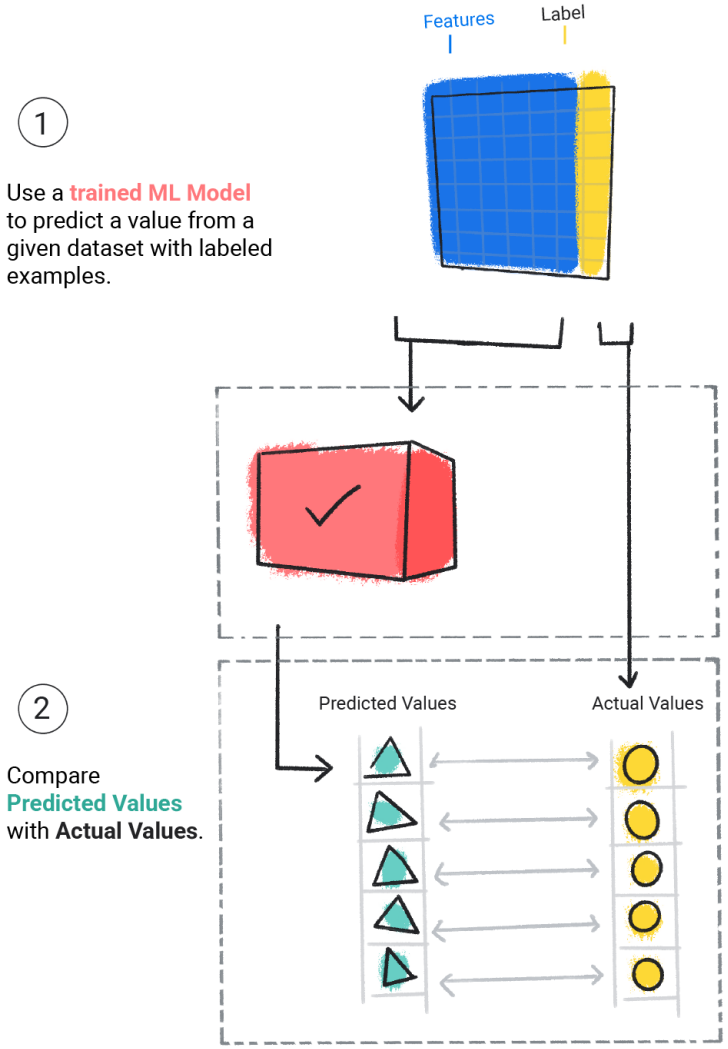
- Regression
- Logistic regression
- Tree-based methods
- Support vector machine
- Neural Networks
- Generative models
- …

We configure

- Parameters
- Features

Evaluation

- Evaluate the performance of models to find the best one

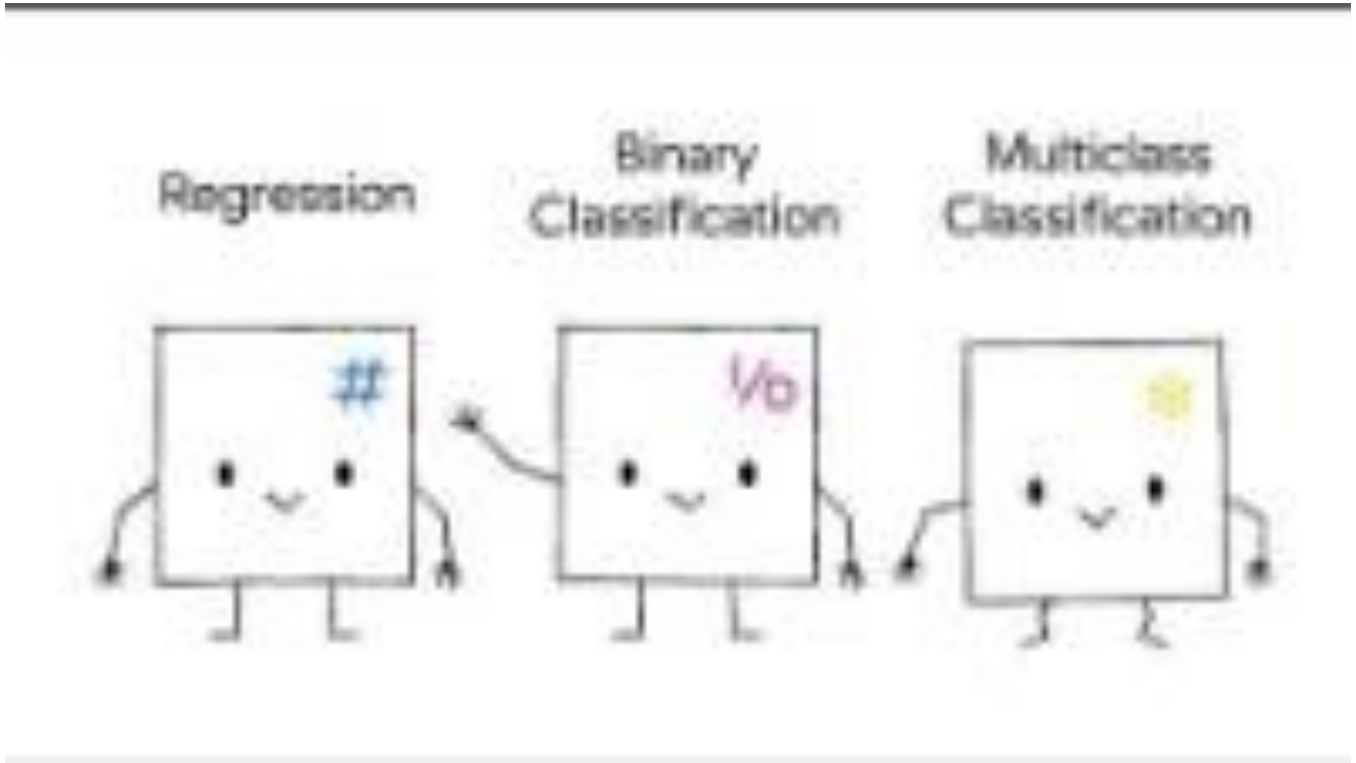


Prediction vs. Inference

- Prediction: main goal is to make accurate prediction
    - E.g., A direct marketing campaign that only cares about to make accurate prediction about which customer will respond
- Inference: to understand how Y changes when X changes
    - E.g., the problem of advertising budget allocation may involve questions like: which media generate the biggest boost in sales? How much increase in sales is associated with a given increase in TV advertising?
- Focus of projects: prediction, inference or both
- Different methods may be appropriate for different focus

Regression vs classification

- In a **regression** problem, Y is quantitative (e.g., continuous variables such as price, wage)
- In a **classification** problem, Y is qualitative, taking values in a finite set (e.g., respond/not respond, win/lose, choice of brands, cancer types)



Unsupervised learning

- No correct answers (labels), just a set of features measured on a set of samples
- Objective is more fuzzy – find groups of samples that behave similarly, features that behave similarly, etc.
- Difficult to know how well you are doing
- The most common use
    - Clustering: cluster data into similar groups based on features

Gene Expression

- The data
  - 64 cell lines
  - Each cell line has 6830 gene expression measurements
  - Associated with 14 cancer types
- We need to
  - Identify whether there are groups, or clusters among the cell lines based on their gene expression measurements