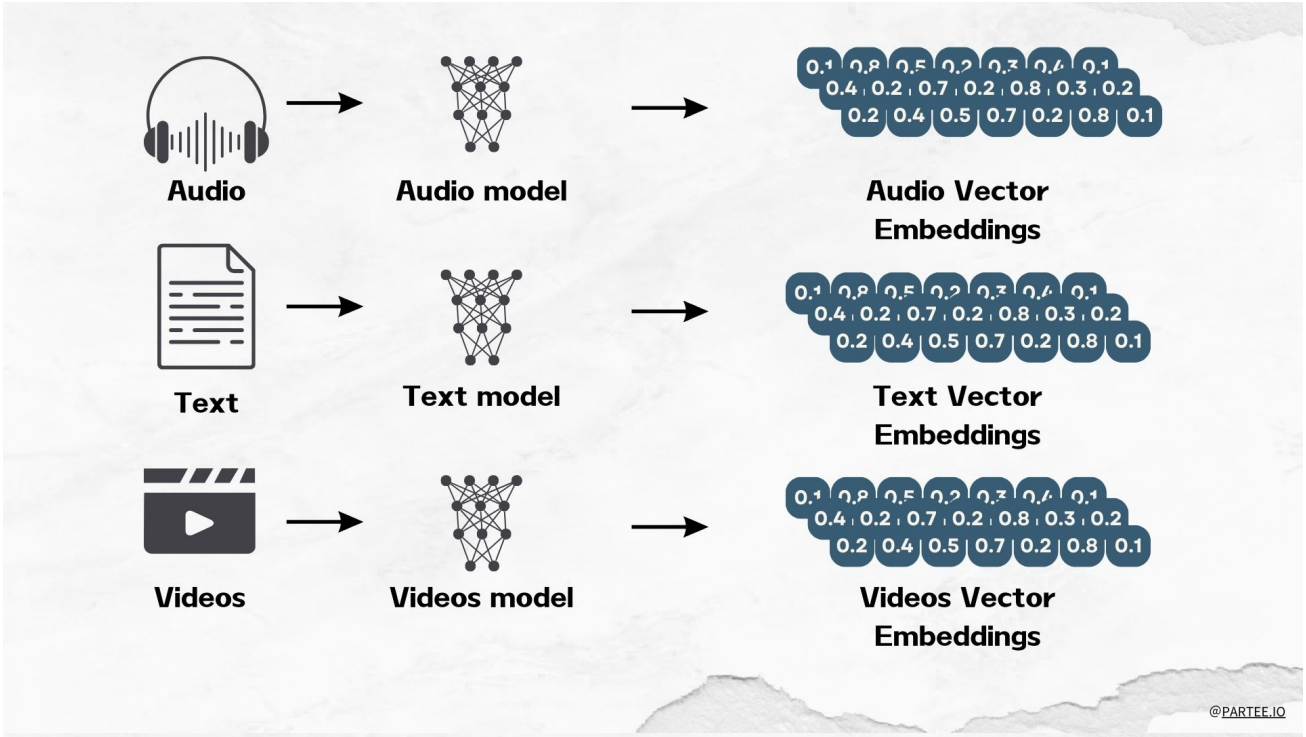
從關鍵字到語義：
探索數據搜尋的演進與向量嵌入技術的突破

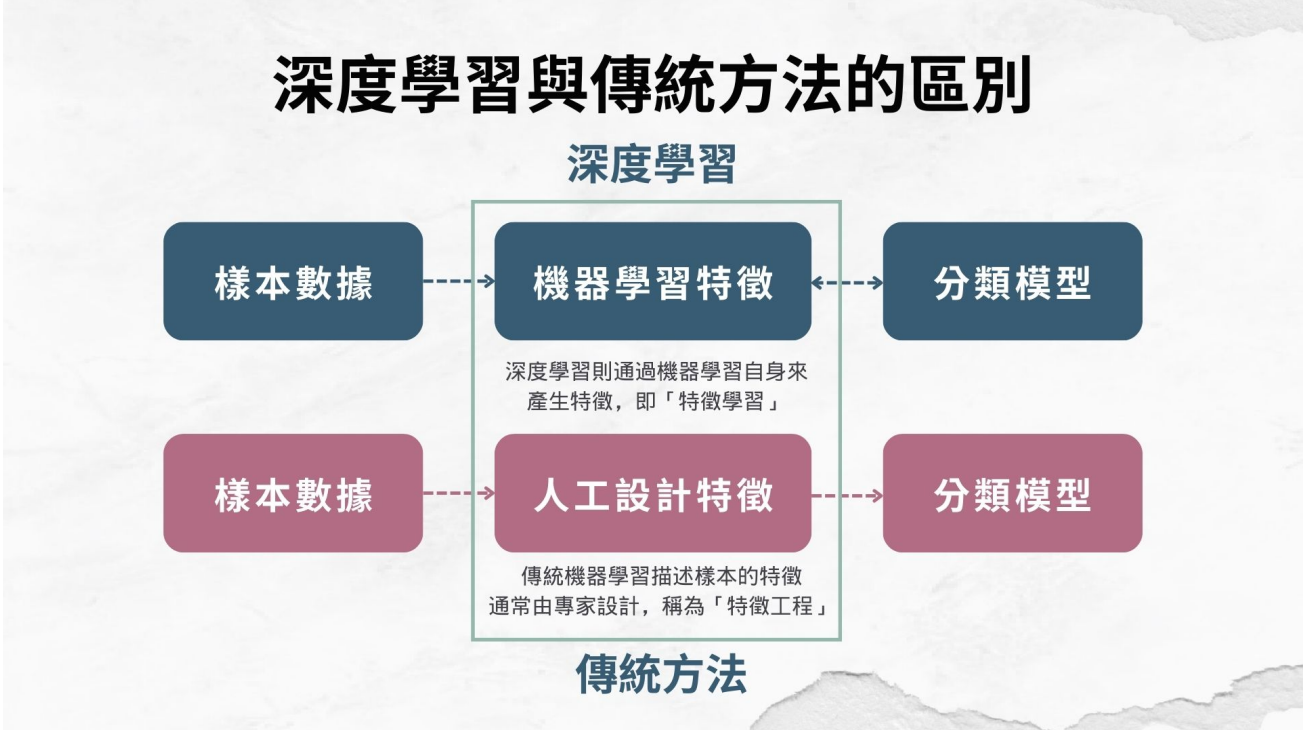
在當代的數據處理領域，傳統資料庫的搜尋機制普遍依賴於特定的索引結構（如 B Tree、倒排索引等）及一系列的精確匹配與排序算法（例如 BM25、TF-IDF）。這種基於文字的精確匹配方法，雖對關鍵字搜尋極為有效，卻在「語義搜尋」的應用上顯示出明顯的局限性。

舉例而言，當進行「小狗」一詞的搜尋時，傳統系統僅能返回含有「小狗」關鍵詞的結果，而無法識別與之語義相近的詞彙如「柯基」、「金毛」等。這是因為這些詞彙在語義上的關聯性超出了傳統系統的識別範疇。於是為了實現語義上的搜尋，開發者不得不人工標註這些詞彙之間的關聯性，通過特徵標籤來建立它們之間的聯繫。這一過程，被稱為 Feature Engineering ，其目的是將原始數據轉化為更能夠體現問題本質的特徵表達。

然而，當面對非結構化數據時，這一方法便顯得力不從心。特別是在處理如圖像、音訊、影片等數據類型時，所需標註的特徵量迅速增加，並且人工標註變得極其困難。在這種背景下，向量嵌入技術（Vector Embedding）應運而生，提供了一種自動化的特徵提取機制。透過 AI 模型（如大型語言模型 LLM），根據特定算法產生的高維向量，能夠代表數據的多維特徵，涵蓋從語彙、語法到語義、情緒等各個層面。

以目前技術為例，文字向量可以通過 OpenAI 的 text-embedding-ada-002 模型生成，圖像向量可以透過 clip-vit-base-patch32 模型生成，而音訊向量則可通過 wav2vec2-base-960h 模型生成。這些由 AI 模型產生的向量蘊含豐富的語義資訊，使得數據的搜尋和分析不再局限於表面的文字匹配，而是能夠深入到語義層面，為語義搜尋帶來了革命性的突破。


Audio
Audio model
Audio Vector Embeddings
Text
Text model
Text Vector Embeddings
Videos
Videos model
Videos Vector Embeddings
@PARTEE.IO


深度學習與傳統方法的區別
深度學習
樣本數據
機器學習特徵
分類模型
深度學習則通過機器學習自身來產生特徵，即「特徵學習」
樣本數據
人工設計特徵
分類模型
傳統機器學習描述樣本的特徵通常由專家設計，稱為「特徵工程」
傳統方法