Simplified explanations

Here are simplified explanations of key terms for the digital sense-making task. The goal is to keep them intuitive and accessible.

- Level 1 is designed for complete beginners in data science, with no need for math background.
- Level 2 is for those who know some basic data science ideas but have little practical experience (high school math is enough).
- Level 3 is for readers who want a deeper dive and are comfortable with technical vocabulary from computer science and data science.

Machine learning

Level 1: Machine learning means teaching computers to find patterns in data. For example, the computer can learn to predict who might have the flu based on their symptoms, or to group customers by their shopping habits.

Level 2: Machine learning is the process where computers discover patterns in data. Sometimes the goal is clear: for instance, classifying images or predicting who has the flu (these are *tasks with answers*). Other times, the goal is to uncover hidden structure, like finding groups of similar customers (these are *tasks without answers*). The outcome of any such process is a trained model, which represents what the computer has learned.

Level 3: Machine learning is the process of discovering patterns in data, broadly divided into different approaches:

- Supervised learning: training on labeled data to predict outcomes, either numbers (regression) or categories (classification).
- Unsupervised learning: working with unlabeled data to find structure, such as groups of similar samples (clustering) or simpler representations (dimensionality reduction).
- Transfer learning: reusing an existing model for a new task, for example, taking a pre-trained image classifier and using it to generate embeddings (numerical descriptions) of images.

Embeddings

Level 1: Embeddings are a way for computers to “understand” images. Since computers work with numbers, each image is turned into a list of numbers. These numbers are chosen so that two similar images get similar lists of numbers.

Level 2: Machine learning needs numbers, but images are pixels, not ready-made numbers for analysis. To solve this, we use *embeddings*: special models that turn each image into a fixed-length vector (a list of numbers). These vectors describe properties of the image, like its content. Two images that look alike will have vectors that are also similar, which means they are close together in a mathematical "feature space."

Level 3: Embeddings are vector representations of data samples. Before applying machine learning, every sample must be expressed numerically. For images, this is challenging because they are high-dimensional and complex. Pre-trained models, such as Inception v3, solve this by learning to classify images (e.g., cat vs. dog). Internally, these models create compact vector representations that capture the essential content of each image. We can extract these vectors—the embeddings—for use in further quantitative analysis, such as clustering or similarity search.

Clustering

Level 1: Clustering means putting similar things together. For example, if we look at students' grades, a computer could group students who score similarly in math or languages. This way, it finds patterns without anyone telling it the groups in advance.

Level 2: Clustering is a type of machine learning task that looks for hidden structure in data without predefined answers. The goal is to group similar items together based on their features. For example, clustering can separate customers with similar buying habits, group medical images with similar characteristics, or organize social media posts by topic.

Level 3: Clustering is an unsupervised learning technique, meaning no target labels are provided. The algorithm tries to uncover the structure of the data by grouping samples that are close to each other in feature space. The number, size, and meaning of clusters depend on both the method (e.g., *k-means*, hierarchical clustering, DBSCAN) and the data itself. Clustering is widely used for tasks such as customer segmentation, medical imaging, exploratory data analysis, and social media analysis.